BUILDING A BETTER TEST WITH CONFIDENCE (TESTING)

By

Paul G. Curran

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Psychology

2012

ABSTRACT

BUILDING A BETTER TEST WITH CONFIDENCE (TESTING)

By

Paul G. Curran

Traditional methods of collecting information from test-takers on multiple choice tests most often involve the dichotomization of individuals into groups based on a simple correct/incorrect criterion. In doing so, information that would further differentiate individuals on ability is lost. This paper shows, though a number of simulations and human subject collections, that this information can be collected and utilized in order to make tests generally more efficient.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Think back to the last time you were administered a multiple choice ability test. Unless you were an expert or an incompetent in the content area of the test you likely engaged in some form of strategic guessing. It is probable that on at least one item you found yourself able to eliminate one or more of the responses, leaving yourself with two or more answers that you found to be appealing choices. One of those responses happens to be the correct answer, and one of two things occurs: 1) you choose the correct answer and are given a score of correct even though you still engaged in what was at root guessing; you are now indistinguishable on this item from an expert who knew the answer readily, or 2) you choose the incorrect answer and are given a score of incorrect even though you had some knowledge about the item; you are now indistinguishable on this item from someone who knew nothing of the topic.

Historically, tests have functioned through precisely this mechanism. Standardized tests have evolved for thousands of years, from ancient China to the modern day (Wainer, 2000). The most common methods of testing rely on assigning individuals a score of correct or incorrect on each individual test item. At the same time, distribution of ability is believed to be better represented as a continuum, such as the number correct score or theta value gained from a sufficiently long test (Lord, 1980).

In modern times there have been a number of attempts to break this habit of dichotomous correct/incorrect collection. Two current examples of note both rely on an item response theory (IRT) framework: the nominal response model (Bock, 1972; Samejima, 1979) and response-time modeled IRT (van der Linden, 2006). Both of these models operate on the principle that information beyond correct/incorrect distinctions exists and can be collected on any given

multiple choice item. One notable flaw of both relates to the fact they are exceptionally complex both in theory and practical use.

A third example of attempts to collect and utilize information beyond correct/incorrect data finds its roots as early as the 1930s. In a 1932 paper, Kate Hevner described a simple addition to a true/false test which produced results both more reliable and more valid than the true/false test given normally. This finding spawned the idea of confidence testing, which persisted for almost 50 years.

Confidence testing operated on a simple idea: the level of confidence that an individual had – or was willing to place – in a response could be used to more finely differentiate individuals on ability. Though this method individuals could be given partial credit on a multiple choice test item instead of traditional correct/incorrect scores. This method flourished for a time, but over the years more and more inconsistent results began to emerge. The benefits of confidence testing appeared almost entirely variable, and without good explanation the research and method slowly faded away.

At its close, confidence testing had become almost unrecognizable from the simple technique put forth by Kate Hevner some half century before. Numerous methods and scoring rules had left the field fragmented with each subset researching and promoting their own versions of the method.

The purpose of this paper is to show that 1) the inconsistent results of confidence testing were relatable to directly measureable test characteristics, 2) all of the different models of confidence testing are in fact specific cases of a more generalized and translatable model, and 3) confidence information can be collected from test-takers in order to improve accuracy of measurement in multiple choice testing beyond simple correct/incorrect methods.

Attempts at Extra Information

As stated above there are, and have been, a number of attempts in the last century to collect information above and beyond correct/incorrect at the item level on multiple choice tests. This idea is not unique to multiple choice items; this problem on other types of items has many clear and intuitive everyday solutions.

For example, anyone who has taught a class and administered a test with open-ended items (e.g. an essay exam or multi-step mathematical problem) knows that students will put in every effort to ensure that partial credit is given on each of those items. Partial credit on open-ended items is in fact quite easy from a mathematical perspective – the practical complexities arise instead in the time and effort required to manually score each item, for each person. Large scale administrations can be prohibitively expensive or time-consuming as to not be viable.

Multiple-choice items are the solution to this time and effort intensive process; they can be administered to enormous populations and scored with relative ease. The cost comes in *the failure to retain the extra information which would have been present if an open-ended item was asked*. While prior attempts have all met with their range of problems there is potential for much to be gained if an adequate solution can be found.

Collecting extra information on multiple choice items has been shown to lead to improvements in test reliability and validity in some cases (e.g. Ebel, 1965; Hevner, 1932; Soderquist, 1936). If these improvements were shown to be consistent this would give the test administrator better precision in estimates of ability, allowing for an improved use as selection or job placement or even class placement tools. An additional and often overlooked benefit is that many of these methods also stand not to *eliminate* guessing, but to allow a valid channel for test-takers to place their choice across multiple answers. In other words, *it gives unsure individuals*

the option to not engage in guessing at all. Guessing is the recourse of a test-taker who is forced to choose one answer over another.

Before looking ahead to possible solutions to this problem it is important to look back at current and prior models which have already attempted this undertaking. The three most notable are those already mentioned. There are two which require the use of IRT: the nominal response model and response-time modeled IRT, and one which does not: confidence testing.

Background: item response theory.

A brief overview of Item Response Theory is necessary before a discussion of either the nominal response model or response-modeled IRT. Put concisely: "Item response theory (IRT) is a general framework for specifying mathematical functions that describe the interactions of persons and test items" (Reckase, 2009, p *v*). These mathematical forms can vary in many ways; it is through the addition, removal, or constraint of parameters that different models are specified.

All (unidimensional) IRT models fit the general idea that the probability of a test-takers' response on an item is a product of a person parameter (θ) and some number of item parameters (e.g. *a, b, c*, etc). Using this information an item characteristic curve (ICC) can be plotted which gives the probability of correct response (for dichotomous items) as a function of ability. This curve will appear slightly different dependent on the number of item parameters used, but the general form is that of a logistic 'S' curve, as seen in figure 1.

Consistent in all models, the person parameter (θ) is simply the trait – often ability in some form – to be measured by the test (Lord, 1980). It is a continuous variable, and the scaling of θ itself is arbitrary. Due to this, θ is commonly scaled to have a mean of zero and a standard deviation of one. The variable θ is the x-axis of the item characteristic curve.

One-parameter IRT models (Lord, 1980) and the Rasch model (1960) use in addition to this person parameter only one item parameter. This parameter (*b*) denotes the difficulty of each item and is in reference to the scaling of the θ-parameter. The point of inflexion of the item characteristic curve occurs when $\theta = b$ (in figure 1; zero). Keeping the scaling of θ constant, the *b*-parameter thus shifts the position of the item characteristic curve laterally, without changing the shape of the curve itself. The larger the *b*-parameter, the harder the item will be.

Two-parameter IRT models (Lord, 1980) incorporate in addition to the *b*-parameter another parameter (*a*) to account for the difference in discrimination between different items. In one-parameter models each curve has the same slope at the point of inflexion; the *a*-parameter allows for this slope to vary. Specifically: "Parameter *a* is proportional to the slope of the curve at the inflexion point [this slope actually is .425**a**(1-*c*)]" (Lord, 1980, p 13). The *c*-parameter will be discussed next, but in two-parameter models is set to zero, reducing the equation to: *slope* = .425 * *a*. As the a-parameter gets larger the curve becomes steeper, reflecting increased discriminability of the item.

Three-parameter IRT models (Lord, 1980) have become some of the most common in use today. In addition to the *a* and *b*-parameters, a third parameter (*c*) is included to account for guessing. The c-parameter can be thought of as the "probability that a person completely lacking in ability ($\theta = -\infty$) will answer the item correctly" (Lord, 1980, p 12). The *c*-parameter is not relevant for items in which the answer cannot be obtained by guessing, as in these cases it drops out of the equation by falling to zero. The c-parameter changes the item characteristic curve by shifting the lower asymptote. In one and two-parameter models the lower bound of the model is 0; it is impossible to have less than a zero percent chance of answering correctly In the three-parameter model this lower bound is simply set to *c* instead of zero.

These four parameters (θ, a, b, and c) form the basis of the majority of IRT models. With this background it is now possible to examine more complex models which incorporate additional parameters.

The nominal response model.

> "To use the binary models, the data have been dichotomized (correct and incorrect) and the distinct identity of the incorrect alternatives has been lost."
>
> (Thissen & Steinberg, 1984, p 501)

The nominal response model was proposed by Bock (1972), extended by Samejima (1979), and detailed by Thissen and Steinberg (1984). The concept behind the nominal response model is that multiple-choice items can be scored for each response choice in order to recapture information that would otherwise be lost to correct/incorrect dichotomization. It was created in order to regain lost information from incorrect answer choices; the idea being that which incorrect answer is chosen has information relating to the ability of the test-taker.

The nominal response model is different from more conventional IRT models in that item parameters exist (and are estimated) at the level of response option. In this way, the number of parameters needed for any given item is dependent on the number of response options for that item (k). Instead of producing only a probability of correct response, the nominal response model produces a probability for each response. In this way the item characteristic curve for every item contains multiple curves equivalent to the number of response options. In the ideal case the correct response produces a logistic curve, while the other curves are more similar to normal curves. A general item characteristic curve produced by the nominal response model (with $k = 4$) can be found in figure 2.

The item parameters in use in the nominal response model are: 1) a vector of parameters defining the discrimination of the response options and thus the slope of the curves (*a*-parameters), 2) a vector of parameters detailing the position of those curves, and 3) a vector of parameters accounting for zero-ability guessing (Thissen & Steinberg, 1984). The first vector is very comparable to conventional *a*-parameters. The second vector of parameters is comparable to the standard difficulty parameters, as it shifts the curves for each response relative to θ. The nomenclature of these parameters varies, but for the purposes of this paper the terminology of γ-parameters will be adopted here. The final vector is used to place individuals who are guessing randomly due to complete lack of ability, and the nomenclature of δ-parameters will be used here. Samejima (1979) originally set this to a vector of constants, but Thissen and Steinberg (1984) make the case that these should instead be estimated.

Unlike conventional IRT models, the nominal response model also requires a number of constraints on these parameters in order to become identifiable. These constraints are on the *a* and γ-parameters in that each set must sum to zero over the different response options within each item. (Thissen & Steinberg, 1984).

In all, this produces a number of free parameters which must be estimated for each item. Thissen and Steinberg (1984, p 503) outline that "the set of free parameters for an item consists of $m_j a_k$'s,…m_j γ_k's,…and (m_j -1) δ_k's,…for a total of 3mj – 1 parameters: 11 for a four-alternative multiple-choice item." [In this terminology, $m_j = k$ = number of response options]. For a five-alternative multiple-choice item, 14 parameters need to be estimated, *for every item on the test*.

The nominal response model has a number of positive characteristics. It was designed to recapture information lost to correct/incorrect dichotomization, and it does. It allows for the fairly powerful ability to model each individual choice on a multiple-choice test item.

This power does not come free, however. Put best by Thissen and Steinberg (1984, p 517, italics added): "The model and its fitting procedures are complex; its use is for 'serious testing,' not classroom exams. We have used sample sizes between *one and two thousand examinees* to estimate the parameters of the model for tests ranging in length from four to 35 items."

The nominal response model has not obtained widespread use because it was not designed for widespread use. This is a point which will be echoed further in this paper: current attempts at collection and utilization of information beyond simple correct/incorrect dichotomization exist at a level of complexity that puts it out of reach of the majority of test administrators. In a selection context this fact alone puts this technique outside of the possibility of the majority of small organizations.

Further, the nominal response model doesn't improve on discrimination of those who have chosen the correct response, *only between those who choose different incorrect responses*. If individuals are predominantly answering correct, there is little left for the model to capitalize on. Tests must also be constructed to have viable and distinct distracter choices, increasing the effort necessary in the test construction phase. If all incorrect items are similar there is nothing gained from learning which one of them an individual has chosen. The effort required to properly utilize the nominal response model is therefore greater at almost all stages of the testing experience, save for administration.

Response-time modeled item response theory.

Response-Time Modeled IRT is based on the idea that the time a test-taker uses to complete an item is a valuable source of extra information for estimating both the ability of the test taker and the characteristics of the test (van der Linden, 2008). "At the level of a fixed person the basic assumption is that the person operates at constant ability and speed. His or her choice of ability and speed level is not free but constrained by a speed-accuracy trade off." (van der Linden, 2006, p 183). In this way the individual can only operate on a test as quickly as their individual ability allows – those with higher ability can achieve the same level of accuracy as a lower ability individual, but quicker.

For the purposes of this paper, an attempt will be made to simplify the description of response-time modeled IRT. Where a conceptual description will suffice it will be favored over one which would be more mathematical. For more in depth mathematics on response-time modeled IRT the author suggests: van der Linden (2006), van der Linden (2007), van der Linden (2008), and van der Linden, Entink, and Fox (2010).

The concept of modeling response time in an IRT model was first proposed in the late 1970s and early 1980s, though recent papers suggest that it has only recently become viable due to the now ubiquitous nature of computerized testing (van der Linden, 2006). Early work on this method was done by Thissen (1983), Scheiblechner (1985), and Roskam (1987), among others, and the general form was to include additional parameters to estimate relating to test-taker speed of response.

There are a number of papers dealing solely with the estimation of individual response time on a test, the first step to incorporating this time into a larger model. Wim van der Linden (2006) has laid out a (relatively) straightforward structure for modeling the pattern of the log of

response time, which includes three parameters very comparable to those found in the two-parameter logistic model.

The three new parameters used by van der Linden (2006) are: α, β, and τ. The τ-parameter is a person parameter, and represents the speed of the test taker's response over the test. It is somewhat similar to θ in other models, at least in use. The β-parameter is an item parameter and represents 'time intensity' or 'time consumingness' of a particular item. The relationship between β and τ can be thought of as comparable to the relationship between the θ and *b*-parameters in the two-parameter logistic model; they are scaled to each other in such a way that the difference between them is what is of actual importance.

The α-parameter can be conceptualized as a type of discrimination parameter. It is defined as "the reciprocal of the standard deviation of the normal distribution" and it is stated that "A larger value for α means less dispersion for the log response time distribution on item *i* for the persons, and hence, better discrimination by the item between distributions of persons with different levels of speed " (van der Linden, 2006, p185).

These parameters can be measured, estimated, and modeled on their own in order to test the fit of this model. This was the impetus of van der Linden (2006), who showed that these parameters fit to a lognormal model of response time with high precision. The next step from this modeling of response time is to model it along-side other parameters of the test and individual. A conceptual hierarchical framework is suggested by van der Linden (2007), which is reprinted here as figure 3.

In this figure, U_{ij} represents the response vector for the test and T_{ij} represents the response-time vector for the test. Each item has *a*, *b*, and *c*-parameters consistent with the three-parameter logistic model, with the addition of α and β-parameters as discussed. Each person has

θ and τ-parameters. The simultaneous estimation of these parameters has been shown to be practically possible, and an empirical study with different subsamples of a sample of 30,000 individuals confirmed that the addition of response-time into the model "tends to improve [parameter estimates'] accuracy" (van der Linden et al., 2010, p344).

Simulation work has also shown that the average improvement in estimation tends to range from 5-20% when the correlation between θ and τ are in the range of .5 to .75 (van der Linden et al., 2010). Prior simulation work has also shown 'substantial' improvement in estimation of ability specifically for the purposes of item selection in computerized adaptive tests (van der Linden, 2008). Work will no doubt continue on the refinement (e.g. van der Linden, 2010; a framework for standardizing the scaling of response-time parameters) and the use of this technique in practical settings.

There are a number of benefits to the utilization of response-time in a testing environment. Due to the fact that this extra information is linked to a behavioral action such as response-time makes it very hard to cheat. As long as a test-taker is attempting to answer the question properly (i.e. not just quickly guessing), the time collected is by definition the minimal time it should take for the test-taker to answer. Additionally, at a very high level this model makes sense in a very simple way; the less you know about something the longer it will take you to recall information about it in order to respond.

At the same time, there are a number of potential downsides that might limit wide-scale use. Just as proponents of the nominal response model claim it is only for serious applications, so too it is likely that the complexity of response-modeled IRT, and the demands of sample size required by the estimation may keep it out of the hands of all but the most skilled and large-scale psychometricians. Response-modeled IRT is not designed to work on a fifth grade weekly

reading quiz, or a custom test designed for employee selection at a small, specialized organization.

While it is almost impossible to *cheat* such a model, it is prone to faking, if only in the incorrect direction. That said, test-takers who get distracted or lose interest may spend more time than necessary on particular items, shifting estimates of their ability. Factors such as test anxiety might also cause test-takers to go slower on a test (Sarason & Stoops, 1978). This test anxiety may even be exacerbated by full disclosure to individuals that their speed on the test will factor into their estimate of ability.

Response-modeled IRT is also – through design – limited to the realm of computer adaptive-like tests. Test-takers must have a response time for each item, and so each item must be presented alone. Test-takers cannot return to items already completed (without large change in test-design). Capturing response time on each item without the aid of a computer, while not impossible, is practically infeasible. This again limits the tests and situations in which this method can be used.

Overall, response-modeled IRT can be considered as a technique that is still in the process of evolution. It is hard to say yet how successful it is or will be, but many factors seem to limit – at the very least – the places where it can be applied.

Confidence testing.

The two models discussed above are relatively modern, and are each predated by a much older field of research which also sought to find methods by which to utilize information beyond a simple correct/incorrect measurement. This field is that of confidence testing, which predates not only the nominal response model and response-time modeled IRT, but item response theory itself.

In the words of Robert Ebel (1965): "In general terms, the examinee is asked to indicate not only what he believes to be the correct answer to a question, but also how certain he is of the correctness of his answer. When his answers are scored he receives more credit for a correct answer given confidently than for one given diffidently. But the penalty for an incorrect answer given confidently is heavy enough to discourage unwarranted pretense of confidence." (p 49).

Confidence (and the very similar concept of subjective probability) have the benefits of being a long-studied area of psychological processes (e.g. Adams, 1961; Baranaski & Petrusic, 1998; Erev, Wallsten, & Budescu, 1994; Koriat, Lichtenstein, & Fischhoff, 1980; McKenzie, Wixted, Noelle, & Gyurjyan, 2001; and Pleskacc & Busemeyer, 2010). Individuals have been shown to be capable of making fairly accurate confidence judgments about their actions in decision tasks (Pleskac & Busemeyer, 2010), and it is through this ability of individuals to be self-aware of their own limitations that confidence testing emerged.

The premise behind confidence testing is that individuals have the capability to report reliably the confidence they have in the answers they give on any given test. The higher the confidence that an individual has in their answer (assuming it is correct), the higher their ability level should be estimated. Theoretically, confidence in a multiple-choice item is constrained across each of the possible answers. The simplest case is that of only two response options; the most common example being a true/false item. This is precisely where confidence testing made its debut. Over time it was extended to items with greater than two response options. Many different methods were built in order to practically implement the constraint between response options, producing a number of different collection methods in use for both items with two response options and items with three or more response options.

The overview which follows is broken into two main sections: first, examining those models built for true/false and two-response items and second, examining those models built for three or more response option items.

Items with two responses (true/false).

Likert-type confidence scale methods.

The first published example of any method of using confidence to provide increased information to test scores was a method described by Kate Hevner in 1932. On her test, Hevner (1932) had test-takers choose one of two responses to each item (in this case true or false). For each item, test-takers also recorded a degree of confidence in that choice, on a scale of one to three. This therefore falls into the category of Likert-type confidence scale collection methods.

Hevner (1932) presented the results of two studies on two different samples. The first sample took only one test, and the second took two tests. These studies tested a number of scoring methods for the data collected from these individuals, including a penalty based scoring of correct minus incorrect similar to a method tested again only a few years later by Soderquist (1936).

The first sample took one test on music, and the best scoring method was a simple weighting of the correct answers based on the confidence ratings. This method produced a Spearman-Brown reliability of .83 reliability compared to .67 for a traditional number correct score. Validity was also tested with correlations to two similar tests. The simple confidence weighted score provided a boost from $r = .41$ to $r = .53$ on the first comparison and $r = .50$ to $r = .70$ on the second.

Hevner's (1932) second sample took two tests using the confidence method. On these tests the boost to Spearman-Brown reliability was from .56 to .70 on the first test and from .66 to

.80 on the second. Validity was only tested on first test (though again with two measures) and was boosted from r = .33 to r = .44 and r = .48 to r = .57.

This 1932 study put forth the basis for a very simple methodology shown to have notable benefits to both reliability and validity. This method was simple enough to be used 80 years ago without any computer aid in administration or scoring. While no formal test-taker perceptions were collected, it was also noted by Hevner (1932) that: "Informal observation among the subjects indicates that the opportunity to express a degree of confidence is a welcome addition to the test, especially when the feeling is one of insecurity." (p 362)

Hevner's (1932) general methods (that is, Likert-type confidence scale on a true-false test) remained without further study until a report of different tests of the method by Robert Ebel in 1965. Ebel (1965) tested this method on a test containing true/false items. Instead of answering true or false on each item, test-takers were provided with five response options ranging from 'This statement is probably true' to 'This statement is probably false.' Each different choice had the chance of producing different score values based on whether the test-taker was on the correct or incorrect side of indifferent.

Correct 'probably true/false' answers earned test-takers 2 points, but at the risk of 2 points deducted if incorrect. Correct 'possibly true/false' answers earned test-takers 1 point at the risk of 0 points deducted if incorrect. Admitting ignorance and answering 'I have no basis for response' earned the test-taker half a point.

In all, Ebel (1965) reported on seven different studies using this method. Three early studies all produced gains in reliability: the first had a gain from .574 with standard testing to .713 with confidence testing, the second .765 with traditional testing to .828 with confidence testing, and the third .728 with traditional testing to .821 with confidence testing. Ebel (1965)

computed an ‘improvement factor’ on these tests based on the Spearman-Brown formula to predict how much a traditional test would need to be lengthened in order to achieve the same reliability as the confidence weighted tests. This ‘improvement factor’ was 1.84, 1.48 and 1.72, respectively.

Following these results Ebel (1965) next reported three other studies conducted on introductory classes in educational measurement from Oct. 1963 to June 1964. Contrary to prior results, these gains to reliability were almost non-existent (.824 traditional, .848 confidence; .790 traditional, .790 confidence; and .8489 traditional to .858 confidence).

Finally, Ebel (1965) attempted another method in July 1964 in which the scoring system was modified based on the idea that the majority of score variance was coming from items answered with most confidence. The score modification was an attempt to get students to only answer those items in which they had the most confidence, and consisted of a scoring system of: Score = 50 + Right – 2*Wrong.

Explained by Ebel (1965) “The uniform base score of 50 represented the expected score from blind guessing and was intended to counteract the naïve feeling that an omitted item necessarily lowers the students score.” The R-2*W part of the formula was intended to encourage a student to omit an item unless he felt the odds favoring a correct answer were better than 2 to 1.” (p 52).

In the end, reliability for this method of confidence weighting produced ‘disastrous’ results; the confidence test had a reliability of .611 against the traditional test’s reliability of .837. Ebel (1965) surmised that the risk of full confidence reduced the spread of scores, producing situations where individuals underrated their own confidence, reducing reliability.

Point allocation confidence methods.

Only a handful of years after the work of Hevner (1932), Harold Soderquist proposed a variant to confidence testing on true/false tests. Soderquist's (1936) method had test-takers choose one of the two responses (true/false) and then allocate points based on their confidence in that response. Test-takers could claim 2, 3,or 4 points on the question with the knowledge that incorrect answers would be deducted twice the points claimed. Claiming 1 point in this case collapsed to traditional scoring and allowed individuals a way to not partake in the experimental condition.

The use of penalty is similar to one of the scoring mechanisms of Hevner (1932), though this was not Hevner's best method (the case with no penalty was). This method may also strike the reader as very similar to the method used by Ebel (1936), and this inherent similarity will be discussed in detail later in the paper when the equations linking each of these methods are derived.

Soderquist's (1936) experimental design tested the difference between scoring the same test either with simple correct minus incorrect or with scores computed based on the points allocated. Soderquist found a boost to reliability from .72 with traditional methods to .85 with confidence methods. This was despite the fact that items scored with allocation of points had less data; participants using point allocation attempted 57.75 answers on average per paper while participants answering normally attempted 74.3 answers on average per paper. This highlights one of the main criticisms that began to appear against confidence testing methods: they take more time and effort of the test-taker than traditional testing alone.

In all, Soderquist (1936) concluded that "The superior reliability shown by the weighted score suggests that weighting for assurance in the true-false test has the same effect as lengthening such a test." (p 291).

Items with three or more responses.

Eliminate believed incorrect response choices.

> "It seems to be a common experience of individuals taking objective tests to feel confident about eliminating some of the wrong alternatives and then guessing from among the remaining ones. This procedure is usually encouraged, as the odds are in the individual's favor."
>
> (Coombs, 1953, p 308)
>
> "It is assumed that partial knowledge exhibits itself in recognizing some of the wrong answers. Complete knowledge, knowing what is right, is equivalent to knowing everything that is wrong. Misinformation is distinguished from partial information in the individual's belief that the right answer is wrong. Objective tests constructed with one right alternative and the others wrong can be administered and scored so as to provide on each item, a scale from complete misinformation through several degrees of partial information to complete information."
>
> (Coombs, 1953, p 308)

The above quotes encapsulate the argument of Coombs (1953), one of the first to extrapolate confidence testing from the two response case to a larger response set. Coombs believed that one way that partial information could be acquired – and partial credit given – was through the confident elimination of incorrect response options. This method is unique in that it only works in cases with three or more response options; in a two response test the confident

elimination of one response option is directly equivalent to simply choosing the other response; no extra information is gained.

Test-takers in Coombs (1953) study were instructed to cross out all response options that they considered incorrect, on the basis that someone fully confident in one response would cross out all others. Test-takers were not required to guess among the choices that they don't cross out; all remaining choices were considered equally weighted. The test was then scored based on two factors: 1) if the correct response was eliminated as incorrect, test-takers received a base score of $1 - k$ (for the remainder of this paper, $k \equiv$ number of response options for any given item), and 2) the number of responses crossed out; each incorrect response crossed out was worth one point. Test-takers therefore had to balance points that could be gained by eliminating more responses against the points that would be lost if a correct response was eliminated. As Coombs (1953) noted, and has been discussed earlier, there is no need for correction for guessing, as it is built into the scoring system.

Unfortunately, Coombs (1953) did not have a large enough sample to make claims about reliability or validity. In fact, results were not reported at all; the paper was framed instead as a proposal for this new method of confidence testing. Coombs (1953) did note that "The first time students are exposed to this method they should be given a clear account of the procedure" (p 310), and that "A small scale tryout at the University of Michigan indicated that the majority of the students were favorably inclined" (p 310).

It should be noted that one of the main limitations stated in the paper has been eliminated in modern times: this technique produced increased labor in scoring tests, something that today can be easily automated with modern computers.

Choose all responses believed correct.

Similar in a way to Coombs (1953) is one of the methods tested by Dressel and Schmid (1953), who examined a number of different methods. Instead of eliminating response options that a test-taker believes are incorrect, in Dressel and Schmid's (1953) 'free-choice' method each test-taker was "…informed that each item had one correct answer, but that they should mark as many choices as needed in order to be sure that they had not omitted the correct answer" (p 578).

Score for this method was calculated as four times the number of correctly marked answers minus the number of marked incorrect answers: 4*(marked correct) – (marked incorrect). Given that items on this test had five response options, a score of zero is produced for an individual who simply marks every response option as correct.

Dressel and Schmid (1953) found that this method of confidence testing produced a small reliability decrement, dropping reliability from .70 on the traditional test to .67 on their 'free-choice' method.

Likert-type confidence scale method.

Dressel and Schmid (1953) also tested a method which they called 'degree of certainty.' This method can be viewed as an extension of Hevner (1932) to the three or more response case; the test-taker was "…directed to indicate how certain he was of the single answer he selected as the correct one by using the certainty scale: 1) Positive, 2) Fairly Certain, 3) Rational Guess, and 4) No defensible basis for choice" (Dressel & Schmid, 1953, p 579).

The score for this method was based on the amount of confidence placed in that option, relative to whether or not it was the correct choice. An individual who chooses the correct answer can gain more points by placing more confidence in that choice at the risk of losing more points if confidence is placed in an incorrect answer. It is important to note that this method had

test-takers only rate one response option on confidence. In this way test-takers could not indicate in any way a second-choice response or rank those remaining. In terms of non-selected responses, those that the test-taker saw as blatantly incorrect were in no way differentiated from those which may have been strong contenders to be selected as correct, in lieu of what was actually chosen.

Dressel and Schmid (1953) found a small reliability benefit of this method, increasing reliability from .70 on the conventional test to .73 on this 'degree of certainty' confidence method. This small difference is comparable in magnitude to Dressel and Schmid's (1953) free-choice method, and may speak to the range of possible effects for their given test and sample characteristics.

Also noteworthy in this method was the fact that individuals were able to complete on average 34.5 'degree of certainty' items over the course of a half hour, very close to the average of 35.2 conventional items that could be completed in the same time. One possible explanation of this result is that the same or similar internal confidence process that is made explicit in the 'degree of certainty' items is occurring but unmeasured in the administration of traditional multiple-choice items.

Jacobs (1971) also tested a method very similar to that of Dressel and Schmid (1953), and in fact more of a direct extension of those techniques originally used by Hevner (1932). Jacobs administered a multiple-choice test on which test-takers both selected their 'most correct answer' as well as their degree of confidence in their response using a three point scale: 'guess,' 'fairly confident,' and 'very confident'.

Jacobs (1971) provided an experimental test in that two different groups were used. The first group was that of a low penalty condition. These students were told that risk and reward

were on a roughly equivalent scale where correct items marked ‘guess/fairly confident/very confident’ would earn 1/2/3 points, respectively, and that incorrect items marked in the same way would cost 0/2/3 points, respectively. In this way the only risk-free choice was an admission of guessing, though in all other situations test-takers only had to risk what they stood to gain. The second group was that of a high penalty condition. For these students, points were earned at the same rate but penalized at 0/4/6 points instead of 0/2/3, twice the risk as could be gained in reward.

In all, Jacobs (1971) found that traditional scoring and confidence scoring in the low penalty group were nearly identical in terms of reliability. Traditional scoring had a higher split-half reliability at .89, compared to the confidence testing split-half reliability of .87. In the high penalty group confidence scoring was drastically worse than traditional testing with a split-half reliability of .39 versus the traditional test split-half reliability of .79.

Multiple correct answers to each item.

As is perhaps becoming evident, Dressel and Schmid tested a number of different methods in their 1953 paper. Two cases which may or may not be strictly classified as confidence testing were those which dealt with multiple-choice questions containing multiple correct answers.

The first of these two examples was the ‘multiple answer test’ in which Dressel and Schmid (1953) “…informed the student that each item might have any number of correct answers, and [that] his score would consist of the number of correctly selected answers minus a correction factor for incorrectly marked answers.” (p 581). Specifically, the “scoring formula used with these items was the number of answers correctly marked minus the marked incorrect answers.” (p 581)

This method achieved the largest boost to reliability of the number of methods tested by Dressel and Schmid (1953), producing a reliability of .78 on the 'multiple answer test' relative to a reliability of .70 on the traditional test, even with the lowest number of items answered of all methods: 28.2 items were completed on average in this method compared to 35.2 items on average in the traditional test.

This method is unique in that items are unconstrained by the common notion of having one correct response embedded in several incorrect responses. A test-taker can no longer find the one correct response and ignore the others; each response must be examined and decided upon in turn. Each response option is no longer constrained by the others, and is in one sense independent of them. An item asked in this fashion cannot rightly be classified as just one item. In truth the test-taker is responding to *k* times as many items as he or she completes (where *k* is the number of response options), as each response option can be considered a dichotomous choice between a correct or incorrect decision by the test-taker. When looked at in this way, test-takers were able to complete 28.2**k* two-response items in this method in the same time that those on the traditional test were able to complete 35.2. Looked at in this way it may be no surprise that reliability would be increased, simply through the (conceptual) addition of items.

Dressel and Schmid's (1953) second method of this type was the 'two answer test' which "…was the conventional test in which one of the incorrect answers was changed to a correct answer so that two out of the five responses were correct. The students were informed of this and also told that their scores would be the number of correct answers" (p 582). This test was thus similar to the 'multiple answer test' except that individuals knew the number of correct response options: two. Removed from this method then is the necessity to make a decision on

each response option. Once two are selected as correct the others are by default classified as incorrect.

Not surprisingly, then, the 'two answer test' did not produce a reliability as high as the 'multiple answer test.' It was scored in a number of ways, the simplest scoring method (the number correct score) producing the largest gains, a boost to a reliability of .76 on the 'two answer test' versus a reliability of .70 on the traditional test.

Point allocation confidence methods.

Michael (1968) based his take on confidence testing in traditional scoring and specifically the common use - in traditional scoring - of correction for guessing factors. As described, the standard correction for guessing used the format of: [correct – (incorrect/(k – 1))]. Michael's (1968) alternative to this concept was 'confidence weighting,' in which "the examinee distributed 10 points among the options for each item and received the total number of points he assigned to the keyed answer" (p 308).

The sample used by Michael (1968) consisted of eleventh and twelfth grade high school students. Test-takers responded to the same test twice; in the first administration they took the test in the traditional manner and were told to choose an answer for every item even if it involved guessing. In the second administration they were given the same test (with their chosen answers still marked) and told to distribute the confidence weights among all answers on each item.

An important note pertaining to the number of points used was that the number of response options (k) was four throughout the test. This made it the case that "the 10-point distribution would not permit examinees to equalize the weights across any given item" (Michael, 1968, p 309). Also, from a mathematical standpoint, "the CW scores were divided by 10 so that a maximum of one point would be available for each item" (Michael, 1968, p 308).

Michael (1968) examined not only how confidence weighting functioned on the entire sample, but also looked at effects on specific sub-groups. On the total sample, confidence weighting provided a boost to split-half reliability from .764 on the traditional form to .840 on the confidence weighting form. This benefit of confidence weighting was slightly higher when both were corrected with the Spearman-Brown formula (.618 traditional to .724 confidence weighting). Michael also noted improvement in a reduction in the standard error of measurement, directly related to the increased reliability.

These benefits of confidence weighting were fairly consistent across high and low IQ groups, with a change in reliability from .639 with traditional testing to .769 with confidence weighting for the high IQ group (.130 increase) and a change from .605 with traditional testing to .723 with confidence weighting in the low IQ group (.118 increase).

On the benefits of this method, Michael "concluded that for a standardized multiple-choice examination in social studies, the CW method affords considerable promise in effecting a higher estimate of [reliability] and a lower [standard error of measurement] than does either the [conventional scoring] or [correction for guessing] method. This conclusion held over a wide range of ability irrespective of sex" (1968, p 311).

Degree of confidence (some derivative of percentile).

None of the methods above represent a capture of the full continuum of possible probability scores that are assumed to underlie a test-taker's responses. It has been admitted in this paper that pure data of this sort may very well be unobtainable without huge leaps forward in measurement and measurement theory. Each prior method has used some different method to categorize test-takers into discrete groups on any item, where the number of those groups is greater than two.

Shuford and Massengill (1967) attempted a more ambitious undertaking in their methods of confidence testing, closer than any other to a true underlying probability/confidence continuum. This ambition, however, seems most focused at monetizing a method of confidence testing. The particulars of their method were proprietary, and only discussed in detail in a piece by Ebel (1968), who was reporting on Shuford and Massengill's (1967) 'Valid Confidence Testing' materials kit.

Their kit contained materials for 5,000 tests at a cost of a little more than $1,100 ($7102.23 in 2010 dollars). Test-takers could select from 26 different degrees of confidence (based on percentages) ranging from 0% to 100% (assumed thus by the author to be increments of 4%). In their method, a student's score is related to the amount of confidence they place in the answer, but not in a linear fashion. Information on some points along the continuum is as follows, Confidence 100%, Weight 1; Confidence 80%, Weight .9; Confidence 60%, Weight .78; Confidence 40%, Weight .6; Confidence 20%, Weight .3; Confidence 0%, Weight -1.

Extrapolation from the information given fails to produce an accurate curve fit to the data using the most common curve estimates; even after removing 0% confidence no exact trend emerges. This would suggest that this method was likely created - at least in part – with an intent for the scoring scale to be difficult to replicate without buying the materials. Ebel makes the claim that "The effect of this [weighting] is to severely penalize dishonest reports of confidence." (1968, p 353).

Unfortunately there are no direct reports of improvement in validity or reliability in either Shuford and Massengill (1967) or Ebel (1968), though authors (Shuford & Massengill) make non-specific claims of improvement. Ebel (1968) further reports indirect evidence that "…indicated degrees of confidence are, at least on some tests, closely related to the proportion of

correct answers given" and that "…Valid Confidence scores correlate substantially but by no means perfectly with conventional choice scores" (1968, p 354).

With 26 response choices for each confidence decision this method can be classified as one of the more complex that has been presented. Discussion of translation between models presented later in this paper will give some idea of the depth of a model this complex. For its complexity, Ebel (1968) reports that on the ease of use: "The developers report that '…students down to the level of the fourth grade can learn to use the materials…'" (p 353-354).

While Shuford and Massengill (1967) and Ebel (1968) did not report exact statistics of Shuford and Massengill's method, Hambleton, Roberts, and Traub (1970) did. This paper reported an empirical study of what they considered a variant of the scoring rules used by Shuford and Massengill (1967). The difference was that Hambleton et al (1970) did not use the same method of 26 possible choices but rather had individuals report percent confidence along an effectively (if not ideally) continuous score sheet.

Hambleton et al's (1970) confidence testing showed reduced split-half reliability: a drop from .710 on the traditional test to .655 on the confidence test. Validity was tested by correlating mid-term exams with final exams, and in this way confidence testing showed improvement: an increase from $r = .621$ on the traditional test to $r = .720$ on the confidence test).

While this seems to be a mixed result and inconsistent with reliability reports by Shuford and Massengill (1967), Hambleton et al (1970) offer an important note of caution: "…the test employed in the study was easy for the group being tested" (p 80) and "What is needed in future studies is a more difficult test" (p 81). This is an idea which will come up in depth further in the paper.

Overall, there are a number of benefits, both empirical and anecdotal, that are associated with confidence testing. Perhaps most overlooked among these benefits are the positive test-taker reactions (Hevner, 1932; Rippey & Voytovich, 1982). Individuals appear to enjoy having the freedom to report partial knowledge in responses. Anyone who has taught or taken a class may recognize this as an outgrowth of individuals' enjoyment of tests which offer partial credit. Perhaps best summarized by Selig (1972, p18, italics added): "How often, when a student looks at a four or five-choice item, *does he want to say to the instructor,* 'I like choice 'd' but there is also a great deal of merit in choice 'a''? He hesitates, tosses a coin, points his pencil, or does whatever students do in such cases and comes up with 'd' or 'a' and marks one of them. It happens to be the wrong choice. As in any testing situation there are two ways to look at the results: the teacher's way and the student's. In this case, the teacher concludes that the student doesn't know anything about the item. But *the student feels cheated* since he did know that one of the two answers was correct."

Confidence testing, at least in its pure forms, offers a large degree of transparency to test takers. In methods such as those involving allocation of points, test-takers know exactly what points they stand to gain if any given response option turns out to be correct. There is face validity to the method relating to the fact that the more willing you are to say that an answer is correct, the more you should know about that answer. And, if nothing else works, individuals have the freedom within the bounds of the model to simply convert any confidence-scored test back into a traditional test by simply always answering with full confidence (as will be shown shortly).

Detractors might suggest that a confidence-scored test will take longer than a traditional test to the degree that the traditional test could have simply been lengthened to provide the same

benefits without the added hassle. Relating to testing speed, Dressel and Schmid (1953) found that: "…students work at different speeds on the various types [of tests] in this order (greatest to slowest rates): conventional, degree-of-certainty [4-point likert on answer chosen], free-choice [choose as many answers to find one correct answer], two-answer [two correct choices to each item, test-takers choose two answers], and multiple-answer [any number of correct response options, choose as many to be certain all correct are chosen]. It seems that students work about 20 per cent fewer multiple-answer items in a given period of thirty minutes than conventional items." (p583-584)

In this case, a reduction of test-taking speed on the most complex method is still only at a rate of about 20%. Over the half hour time limit the difference between conventional testing and degree-of-certainty (the next fastest method) was less than one item: those in the conventional condition completed 35.2 items on average, and those in degree-of-certainty completed 34.5 on average. This relates to a loss of about three items for every two hours of testing, which is hardly arguable as anything but negligible.

Another problem relating to confidence testing was the failure over time to ever standardize a singular technique, or lay down formulas for the translation of confidences obtained in one method to confidences obtained in another. Echternacht (1972) appears to have been the only individual to attempt anything related to this idea; he translated fixed point allocations to probabilities. Individuals given five points (in this case, stars) to allocate across five response options were assumed to have probabilities of correct response corresponding to the number of points (stars) allocated to the correct response divided by five. While this is certainly better than nothing, it can hardly be considered unifying.

The lack of this standardization is not a problem inherent with confidence testing itself. This paper will argue and demonstrate that mathematics for translation between methods can be produced with minimal assumptions.

Confidence testing suffered from an inability to produce not only consistent gains, but simply consistent results over time. Confidence testing was shown to produce nearly the entire range from detriment to benefit. This paper will use simulated pilot data in order to argue and demonstrate that these differences in results are likely attributable to unexplored (and mostly unreported) test characteristics.

Finally, it has been suggested that the proclivity of an individual to use confidence judgments rather than assigning complete confidence to a single response (inconsistent with their actual confidence) could be related to individual difference personality factors of the individual (Hansen, 1971). Hanson reported that "Individuals who indicate a preference for risky options…tend to be more certain in their responses than would be typical for an individual with their knowledge" (Hansen, 1971, *p* 13). In addition, it was also shown that risk-taking was *not* correlated with test scores themselves (r = .043, -.025, *ns*). This, then, is not as disastrous a result as might be expected. It does mean that if confidence testing is producing more reliable and more valid tests, that some individuals will have more reliable and more valid scores than other individuals on the same test. However, those risk-taking individuals (guessers) who have lower reliability and validity will still have the reliability and validity that they would have had on a traditional test. Those who are not engaging in risk-taking (guessing) will gain the benefits of a more reliable and valid test. To make a practical example: in an adaptive test this would mean that risk-takers (guessers) would have to take more items to obtain the same level of reliability and validity as those who were not risk-taking.

Hansen (1971) did not find correlations with the use of confidence testing and other personality factors, such as test anxiety, though it is difficult to say that they do not have any relationship. Due to the haphazard nature of confidence testing research at the time, even Hansen's (1971) risk-taking findings must be taken with some skepticism. If the use and benefit of confidence testing is shown to be related to characteristics of the test that were unmeasured (or at least unreported) at the time of Hansen's results, then replication and construction of individual difference linkages from the ground up is necessary.

That is not to say that prior work is completely without merit. Hansen's (1971) examination of risk-taking and test anxiety are a good starting point for examination of individual differences as they relate to confidence testing. In fact, years earlier, anecdotal findings of Hevner (1932) noted that "Informal observation among the subjects indicates that the opportunity to express a degree of confidence is a welcome addition to the test, *especially when the feeling is one of insecurity*" (*p* 362, italics added). This general insecurity on a test has links to the anxiety aspect of Kanfer and Heggestad's (1997) motivational taxonomy, and also somewhat to general avoidance motivation (e.g. Elliot, 1999).

With only questionable prior research on how these different individual differences might relate to confidence testing, we are left with somewhat of a thought experiment. Kanfer and Heggestad (1997) list a number of individual difference traits which fall under the overarching theme of anxiety. Arguably most relevant to this current study are the concepts of general anxiety and specific test anxiety (as already argued for by Hansen, 1971). Presented with a confidence scored test, the question now is how individuals high or low on these behaviors should act.

Hansen's (1971) results for risk-taking seem perfectly reasonable – individuals who score high on risk-taking are more likely to 'put all their eggs in one basket', so to speak. They are more willing to take the risk that the answer they are most confident in is the correct answer, thereby placing all confidence in that answer. Individuals who are less risk-taking are more likely to distribute their confidence accurately, hedging against the possibility that their most confident choice might still be incorrect.

The prospects for anxiety (both general and specific) are also consistent with reasonable expectation. While not significant, Hansen (1971) found a negative relationship between test anxiety and the amount of confidence individuals were willing to place in singular answers. This means that those individuals high on test anxiety were more likely to use confidence testing, while those low on test anxiety were less likely to use it. This relates back to the idea of anxiety and avoidance, and even risk-taking. In adopting an avoidance oriented stance regarding the test, they should be less likely to take risks by displaying overconfidence. Those who are anxious about testing (or perhaps just anxious in general) should be more diffident about answering with full confidence, and may be more likely to distribute their confidence in a way which guarantees them at least some credit.

These individual difference measures, along with others which will be discussed later in the paper, will be collected in order to shed light on how some of these initial constructs of interest may influence confidence testing.

Integrated General Form of Confidence Testing Model

It appears no overstatement to represent classical confidence testing as a fragmented field. Numerous methods for the collection and utilization of this data have been presented, and for each method which was published there was likely another which was lost to time. Indeed, one of the problems in the field of confidence testing was that a wide range of different techniques were all pitted against one another in an attempt to find which worked best. Stakes were relatively high for those who were looking to sell their method, as can be seen in the case of Shuford and Massengill (1967).

What was never discussed, and perhaps never examined, was the fact that each of these methods are in fact collapsed specific cases of a more generalized model, at least from a mathematical standpoint. Using the assumptions already made by the creators of these different models, a generalized model can be created which provides a linkage between them. Such will be the purpose of this section. A mathematical linkage between methods will allow for tests of differences that should arise from psychological processes; if methods are mathematically identical then psychological processes are all that remain in accounting for differences. The confound of the mathematical framework of the different methods can thus be controlled for in analysis. Defining the mathematics of these models opens an entirely new way to do research on and pertaining to them.

The basic idea that holds across all models of confidence testing is that weights are given to each response option through some translation of collected confidence information into score. At the most basic the score on the item is the amount of weight a test-taker allocates to the correct response, however that may be scaled. As discussed, the ideal (and believed to be improbable) case is that where information is collected relating to the actual probability that a

test-taker would choose each different response option. In this ideal case the probability of response maps to the weight each answer has in their choice process. Each method of confidence testing uses a different method to try to approximate this continuous confidence/probability scale.

In the case of two response options (r_1 and r_2, most often true/false), information need only be collected on the response option chosen. Confidence on response one (Cr_1) and confidence on response two (Cr_2) must sum to one, which means that: $Cr_2 = 1 - Cr_1$ and $Cr_1 = 1 - Cr_2$. This is not only a property of items with two responses, and in fact it is simply a special case of a constraint which takes place on all items. In all methods of confidence collection examined so far, it is the case that information must be collected explicitly or implicitly on k - 1 of those response options in order to have full confidence information about that item. The weighting on the final response is constrained by the responses to the other response options and can thus be implied from them. Mathematically, once an individual has supplied weights to k - 1 response options the final response option is constrained to be equal to $1 - (\sum (Cr_1, Cr_2 \ldots Cr_{k-1}))$.

In the case where an individual has no knowledge whatsoever concerning the answer to an item (e.g. taking a test in a completely foreign language) the probability of answering correctly can be defined as equivalent to $1/k$, which may be recognizable as the starting value in the estimation of the guessing parameter. This probability is not only the probability of choosing correctly, but rather of choosing any response – each of the response options is weighted equally at $1/k$.

Consider now Coombs (1953) method of eliminating all answers confidently known to be wrong. In this method the individual is simply reducing k through the process of elimination,

where the probability of choosing any of the remaining options is now assumed to be $1/(k - e)$, where e is defined as the number of response options eliminated.

Frederic Lord himself considered this possibility in his work on item response theory: "We might next imagine examinees who have just enough ability to eliminate one (or two, or three…) of the incorrect alternatives from consideration, although still lacking any knowledge of the correct answer. Such examinees might be expected to have a chance of 1/(A-1) (or 1/(A-2), 1/(A-3)…) of answering the item correctly, perhaps producing an item response function looking like a staircase." (Lord, 1980, p 17).

Similar to Coombs (1953) is Dressel and Schmid's (1953) method of choosing as many response options as it takes to be confident that one of them is correct. It is easy to see that this is simply a different presentation of Coombs (1953) basic idea; those answers not chosen in the group of 'believed correct' have essentially been eliminated. We can now define c as the number of response options chosen in the group 'believed correct', and equivalent to $k - e$. Because $c = k - e$ can be rewritten to show that $e = k - c$, the probability of choosing any of those answers in the group of 'believed correct' can therefore be shown to be equal to $1/(k - (k - c))$ or simply $1/c$.

The allocation of points method (Michael, 1968), fits this model in that the number of points allocated to any given response can be shown to fit the degree of probability of choosing that response. The confidence weight (Cr_0) for each response can be set equal to $P*(1/p)$ where P is the number of points allocated to that response and p is the number of points allocated overall. This is consistent with Echternact (1972), who computed probabilities with P*(1/5) on a test, where p was fixed at 5. In actual use, the number of points allocated is not bounded in any way. In fact, p and P can be set to any positive real integer, from one to infinity (P can

additionally be set to zero). What is important is the *proportion* of points allocated to the correct answer relative to the collection of points allocated on that item. Interestingly, this point-based model can be shown to include, as a special case, the models above involving elimination and inclusion. In these cases P is simply fixed at 1 and $(k - e) = c = p$.

Further, traditional testing is also a specific case of this model. Test-takers are given one point to be allocated across all response options; in this case p is fixed at 1 and P is constrained to either 1 or 0 for each response. The equation then collapses to 1/1 or 0/1, or rather 1 and 0: correct or incorrect. In this way, a person taking a confidence scored test can, at-will, turn it into a traditional testing situation.

The use of Likert-type scales of confidence also fit a specific case of this model, though a bit more math is required. The simplest case is, again, the case of items with two response options and was the basis for the entire range of models to follow as introduced in Hevner (1932). Hevner (1932) had test-takers report confidence in their response on a three point ordinal scale. If we are assuming that this scale is truly Likert-type (as we have been throughout the paper) then with this distinction this scale also gains the concept of interval measurement. The movement from one level of confidence to the next is assumed to be constant across all levels of confidence. This provides almost all that is needed for deducing the underlying mathematical framework.

Before this, however, another very important point must be made. Due to the way Hevner (1932) anchored the Likert-type scale there are actually five weights that can be implied from these data. Confidence choices were, put simply, ‘very sure’, ‘fairly sure’, and ‘not at all sure’; thus a test-taker could be ‘very sure’, ‘fairly sure’ or ‘not at all sure in choice A and ‘very sure’, ‘fairly sure’, or ‘not at all sure’ in choice B.

As discussed earlier, the ‘not at all sure’ is an admission of guessing and weights each response with a half probability of choice; .50 if we follow the convention that all confidence weights should sum to 1. Therefore, the weight for ‘not at all sure’ is identical regardless of whether it is in choice A or choice B – they need not even be marked in this case.

A full five-point scale is the result: ‘very sure in choice A’, ‘fairly sure in choice A’, ‘not at all sure’, ‘fairly sure in choice B’, and ‘very sure in choice B’. Just as before, implied responses can be made given that $k - 1$ confidence ratings are collected. This produces an implied symmetric reverse scale parallel to this five point scale as follows.

Given response of ‘Very sure that choice A is *correct*’ implies that ‘Very sure that choice B is *incorrect*’ and weights A at 1 and B at 0. Given response of ‘Fairly sure that choice A is *correct*’ implies that ‘Fairly sure that choice B is *incorrect*’ and weights A at .75 and B at .25. Given response of ‘Not sure at all (in either)’ implies that ‘Not sure at all (in other)’ and weights A at .50 and B at .50. Given response of ‘Fairly sure that choice B is *correct*’ implies that ‘Fairly sure that choice A is *incorrect*’ and weights A at .25 and B at .75. Given response of ‘Very sure that choice B is *correct*’ implies that ‘Very sure that choice A is *incorrect*’ and weights A at 0 and B at 1.

Weights are calculated through a translation into the point allocation method calculated above. For any Likert-style scale where L is the number of collected Likert scale points, p can be shown to be equal to $(k*(L - 1))$, in this case $p = 4$. The implication of this is that allocation of $(k*(L - 1))$ points can replicate an L-point Likert-style confidence scale for an item any number of response options.

These points allocated to each answer are simply multiples from the above example, and can be calculated by multiplying each of the above allocations (0, .25, .5, .75, and 1) by 4 to come up with the interger points that would be allocated.

An individual who is 'very sure' that choice A is correct will allocate all four points to choice A and zero to choice B. Being 'fairly sure' a participant can hedge their bets and allocate three points to choice A and one to choice B. An admission of guessing comes from the split allocation of two points to each choice. To note is the fact that in the two-response special case it does not make sense to consider 'choosing' choice A and then allocating only one point to it. The test-taker instead should have chosen choice B. While these options (e.g. allocating 1 point to choice A) are not directly available to an individual (at least practically) they are still part of the scale as they can be implied. In the extension to cases of three or more response options these situations do become much more practically and explicitly available.

The simplest Likert-style confidence method involving three or more response options is that of Dressel and Schmid (1953). Test-takers were "…directed to indicate how certain he was of the single answer he selected as the correct one by using the certainty scale: 1) Positive, 2) Fairly Certain, 3) Rational Guess, and 4) No defensible basis for choice." (p 579)

Still assuming interval scale qualities but for simplicity and space-saving measures, we can cut this down to three confidence options similar to Hevner's (1932) scale: positive or 'very sure', 'fairly sure', and 'pure guess'. One will have to accept on faith (or complete the proof themselves) that an extension of the math can be shown to work for any number of confidence options.

This scale (Dressel and Schmid, 1953) was only collected on the answer chosen correct and again only represents a portion of the full implied scale. Weights can be crafted for the

answer chosen in the same way as was done to the Hevner (1932) scale. However, without collection of confidence on *k* -1 responses, assumptions must be made regarding the spread of effectively 'leftover' weighting. If an individual is 'fairly certain' of response A what does that mean for the remaining options?

The choice of 'pure guess' represents a weight of 1/*k* in all response options; in the case of *k* = 2 this broke down to a 50-50 chance. This effectively sets a floor for which confidence in a response, if chosen as correct, should never drop below. In the case of *k* = 3, the response of 'pure guess' instead results in .33 in all response options. It then follows that the other more confident responses should cover the range of the scale from that point (.33) to 1.

This information can be found in table 1.

The number of points that need to be allocated to match this scale is found by k*(L-1); in this case six. However, this comes with some caveats due to the constraints of giving confidence only in the response which has been chosen.

This information can be found in table 2.

This caveat is that each step up the Likert scale from 'pure guess' to 'very sure' results in an increase in allocation of (*k* - 1) points (in this case two). This is due to the fact that other points need to be allocated to all other responses, at least assuming that points *in this fashion* can only be distributed as whole integers.

To extrapolate to different situations, then, if there were four response options (*k* = 4) this method would be equivalent to allocating 8 points [(4*(3-1))] in steps of 3 ('very sure' = 8 points, 'fairly sure' = 5 points, 'pure guess' = 2 points). The case of 8 points is an allocation of all to one response, 5 points is an allocation of 5 to one and 1 to each of the other 3, and an allocation of 2 gives 2 to each of the four options.

To go back to Dressel and Schmid (1953), their method with a four point Likert scale can be shown to actually be representing the allocation of 9 points if $k = 3$ [(3*(4-1)], 12 points if $k = 4$ [(4*(4-1))], or 15 points if $k = 5$ [5*(4-1)]. The actual number of response options on their test (k) was unable to be discovered. However, an allocation of 15 points is equivalent to probabilities starting at 0 and incrementing to 1 in steps of .065 – much greater resolution than one would expect to be able to collect if asked from a straight probability standpoint. Worth noting is the fact that this is *still not as resolute of a scale* as that proposed by Shuford and Massengill (1967), who incremented from 0 to 1 in steps of .040.

In all, the general form then follows that the weight for any given response can be defined as:

$$Weight \equiv P_A * 1/p$$

Where P_A is the number of points allocated to that response, and p is the number of points allocated overall.

For the representation of the method of exclusion of incorrect answers:

$p = k - e$ and $P_A = 1$

Where e is the number of responses excluded and k is the number of response options on each item.

For the representation of the method of inclusion of correct answers:

$p = c$ and $P_A = 1$

Where c is the number of response options marked correct.

For the representation of the method of a traditional test:

$p = 1$

For the representation of Likert confidence ratings:

$p = k^*(L_{MAX} - 1)$ and $P_A = (L_0/L_{MAX}) * p$

Where L_0 is the ascending rank order number of the Likert rating chosen and L_{MAX} is the number of Likert ratings available; the lowest representing pure guessing.

This generalized model thereby unifies the highly disparate methods used by the field of confidence testing. Specific forms of the model can thus be translated mathematically from one model to another, and compared. Differences that occur in practical use between two mathematically equated models infer that differences are not mathematical but psychological. Psychological factors can thus be more purely studied without mathematical confounds. It is also the case that one form may be found to be easier to understand or more preferable when compared against other mathematically equivalent forms.

Additionally, just as traditional testing is nested in this model so are all the different specific methods. With an overarching framework there is actually no need to constrain a test to one method or another. In the same way that a test-taker can 'break' confidence testing by always putting confidence in one answer (thus making it a traditional test), different methods can be used at-will. With a mathematical framework by which to calculate comparable weights across methods individuals can use whatever method they chose – *at the item level.*

Consider a multiple-choice item with four response options. A test-taker may only be able to eliminate one answer that they know is incorrect. This is the exclusion of incorrect method (or a 3-point allocation method) and would set weights at 0/.33/.33./.33.

On the next item the test-taker may decide that two of the answers seem correct, and that they can't decide between them. This is the inclusion method (or a 2-point allocation method) and would set weights at 0/0/.5/.5.

On another item the test-taker has narrowed the choice to two answers, but feels slightly better about one of them. They may choose to allocate 3-points shared between the two responses with one of the responses getting 2-points and the other getting only 1. This would set weights at 0/0/.33/.66.

On another item the test-taker may have full confidence in their answer and eliminate all other answers, simply select that choice, or allocate N points to it (where N is any real integer), and none to any other answer. All of these scenarios collapse the item to that found on a traditional test, producing a weighting of 0/0/0/1.

It is thus the case that in the worst case situation all test-takers collapse all items to traditional items, producing a test that is as reliable and as valid as it normally would have been. In such a situation, all that is lost is the time training the test-takers.

Pilot Hypotheses

There are a number of mathematical hypotheses following from above arguments that can be examined using simulated data. First and foremost:

H_P1: Confidence testing will produce a more reliable test than an equivalent traditional test

H_P2: Confidence testing will produce a more valid test than an equivalent traditional test

On the differences between different resolutions of similar methods:

H_P3:Benefits of confidence testing will be larger the closer the method approaches raw probability data (e.g. 10-point allocation better than 5-point, 5-point better than 2-point, etc)

To borrow from Dressel and Schmid, (1953, p 576) it is the accepted belief that "…the student whose response contains an element of guessing will tend to miss enough items over an entire test to differentiate him from the student who responds with complete certainty." That is, given enough items on a test, a traditional test should begin to converge on true score. In this way:

H_P4: Benefits of confidence testing will be related to test length in that shorter tests will produce larger <u>gains</u> than longer tests.

To borrow from Echternacht (1972, p224, italics added), summarizing De Finetti (1965): "If the examinee were certain of the correct answer, the best response was just that, and the problem disappears; but oftentimes, *especially if the item was difficult*, the examinee had a degree of uncertainty about his action." That is, less guessing occurs the easier a test becomes, as test-takers are fully confident in their responses. On more difficult tests confidence testing is able to capitalize on information that is lost to what would become guessing. In this way:

H_P5: Benefits of confidence testing will be related to test difficulty in that more difficult tests will produce larger <u>gains</u> than easier tests.

Finally, the discrimination of the items on a test should have an impact on the benefits of confidence testing. This should be due to the fact that items which have high discrimination are able to place individuals accurately into correct and incorrect with few individuals falling into

the probabilistic space between (where confidence testing finds its extra information). In this way:

H_P6: Benefits of confidence testing will be related to test discrimination in that tests with lower discrimination will produce larger <u>gains</u> than tests with high discrimination.

Pilot Studies

Pilot Simulation #1 – Comparison of Different Confidence Testing Methods

Pilot data was simulated in order to illustrate the fact that useful information is lost about individuals when they are artificially dichotomized on each item into categories of correct/incorrect. Data was simulated for 200 individuals on a 60 item test using the following method:

1) Individuals were sampled from a normal distribution on ability (θ: mean = 0, SD = 1).
2) The test was constructed to be of average difficulty relative to the sample (*b*-parameter: mean = 0, SD = 1), average discriminability (*a*-parameter: drawn from lognormal with parameters -.13, .34), and a chance of guessing as might be expected on a multiple choice test with 5 response options (*c*-parameter: mean = .2, SD = .1).
3) Probability of correct response was computed using the 3-parameter logistic model of the form:

$$P(\theta) = c + (1-c) * \frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}$$

4) Uniform random data between 0 and 1 was generated for each item of each individual. These data were then used to create dichotomous correct/incorrect information for each item of each individual through a comparison process. If the generated random number was less than their probability of getting the item correct, they were given a score of

correct (1). If the generated random number was higher than their probability of getting the item correct, they were given a score of incorrect (0). These data were then summed at the individual level to give each individual a total correct score for the test ranging from 0 to 60. This will be considered the control condition of standard or traditional testing.

The ideal case of confidence testing assumes that individuals are capable of self-reporting confidences or subjective probabilities that correlate perfectly with their actual probability of getting an item correct. Raw probability of correct response for each item for each individual as computed above could be used as the ideal confidence testing collection, though this ideal case is admitted to be impossible from a practical standpoint. Instead, these probabilities can be used as input into methods which are more likely to approximate data which can be practically collected from an actual human sample.

In order to degrade the ideal case allowing for the fact that individuals are likely to be unable to give their raw probability with exact precision, a number of additional possible outcome matrixes were created in similar fashion. These scores represent *the share of 100% confidence that they are likely to give to the actual correct answer*. What remains (e.g. [100% - score] is that confidence which would have been given to incorrect answers). Each method can also produce a "total correct score": a summation of the probabilities of correct answers for each individual. These different methods are described using the 'allocation of points' method due to the fact that the generalized form above can be used to translate any other method to and from this framework.

1) 10-point allocation (Michael, 1968). These data replicate a situation where individuals have 10 points to distribute across all the response choices for a given item. This creates 11 possible confidences for each item: 0%, 10%, 20%, 30%, 40%...90%, 100%. This pattern was simulated by rounding each raw probability (represented from 0 to 1) to the nearest tenths place, producing the above confidences.
2) 5-point allocation (Ebel, 1965). These data replicate a situation where individuals have 5 points to distribute across all the response choices for a given item. This creates 6 possible confidences for each item: 0%, 20%, 40%, 60%, 80%, 100%. This pattern was simulated by rounding each raw probability to the nearest of these values.
3) 3-point allocation (similar to Hevner, 1932). These data replicate a situation where individuals have only 3 points to distribute across all the response choices for a given item. This creates 4 possible confidences for each item: 0%, 33%, 66%, 100%. This pattern was simulated by rounding each raw probability to the nearest of these values.
4) 2-point allocation. These data replicate a response situation where individuals have only 2 points to distribute across all the response choices for a given item. This creates 3 possible confidences for each item: 0%, 50%, and 100%. This pattern was simulated by rounding each raw probability to the nearest of these values.

For each simulated set of data a number of statistics were calculated. A benefit of simulated data is the fact that the "true scores" of individuals are known, as they were used to generate the data. Thus, correlations between the "total correct score" for each test and the original ability (θ) were calculated, as well as the resultant *r*-square values. This allows for *r*-square change comparisons relative to the original standard testing model. Further, reliability

was computed on the raw item-by-individual matrix of data for each different simulated method. These results are presented in table 3 below.

These results show that from a mathematical standpoint there is extra information above correct and incorrect contained not only in the raw probability scores but in several degraded forms of the data. In terms of relationship with true score, an *r*-square increase of .0715 and an increase in reliability of .086 can be gained from even finding confidence simply to the degree of 0/.33/.66/1. Further, even a collection of one more piece of information above standard – finding when individuals are torn in decision between two responses (i.e. a confidence collection of 0/.5/1) – produces an almost identical r-square increase of .0692 and reliability increase of .081.

Pilot Simulation #2 – Introduction of Error; Comparison to Traditional Testing

The results of pilot data simulation #1 show that there does not appear to be appreciable difference between the different number of breaks made to the confidence scale. That is, the trouble of a 10-point allocation (or even of the ideal case of raw probabilities) may not hold much more benefit than simpler methods. This allows for a reduction in the number of methods to be simulated as the simulations become more complex. For the following simulation only the 5-point and 2-point allocation methods will be simulated.

The data of the first simulation makes an assumption about the ability of individuals to effectively round their raw probability score to one of the adjacent confidence levels. This was somewhat admissible in the first pilot as comparison was being made between different methods of confidence collection. To fairly compare methods of confidence collection to traditional testing the same random error that is being introduced into the simulated traditional data must also be introduced into the simulated confidence data.

In order to incorporate the same error into the confidence data, the same random number matrix that was used to create the traditional scores was used. Recall that this random number matrix was created with a number between .00 and 1.00 for every person and every item. If their probability of obtaining a correct score was greater than this random number they received a score of correct, if their probability was less than this random number they received a score of incorrect.

The confidence data can be conceptualized piecewise in a very similar way. For example, in the case of 2-point allocation a person can have three different levels of confidence in the right answer: 0%, 50%, or 100%. This comes from the fact that they can allocate no points to the correct answer (0%), one point to the correct answer (50%), or both points to the correct answer (100%).

The simulated data provides a raw probability of correct response which in the ideal case would be the confidence that an individual places in the correct response. In pilot #1 this probability was rounded to the nearest possible level. For example, an individual who had a .30 probability of correct response would be rounded to .50 in the 2-point allocation model. Instead of simply rounding, these generated random numbers can be used in the same fashion as in the traditional simulation.

Just as an individual has a probability of getting a score of correct or incorrect in a traditional test, so too should an individual have a probability of obtaining the higher or lower confidence in their answer based on their raw probability. To continue to use the example of .30 probability on a 2-point allocation, the individual should not be simply moved to the nearest level, but placed there probabilistically.

Conceptually there are a number of ways to envision this process, all mathematically equivalent. In essence a random number can be drawn from the range of .00 to .50 for comparison and placement – if the probability is higher the individual will be placed at .50, and if the probability is lower than the individual will be placed at .00. The individual with original probability of .30 now has a 60% chance of obtaining a .50 confidence and a 40% chance of obtaining a .00. Another possibility is for the probability scale from (in this case) .00 to .50 to be scaled to represent the full .00 to 1.00 scale. This method is also one in which the random number matrix of .00 to 1.00 that has already been generated can be used directly. In this way *the noise that is introduced is exactly identical to that which has been introduced into the simulation of the traditional test.*

Using these results produces confidence scores which can be more fairly compared to those of the traditional test. Results based on these scores can be found in table 4.

These results show that even the addition of random noise similar to that used in simulating the traditional test results does not noticeably reduce the possible gains of a confidence testing methodology. Further, similar to the results of pilot #1, there does not seem to be drastic difference between the benefits of 5-point and 2-point allocation.

Pilot Simulation #3 – Effect of Test Length

While the findings so far are promising, a discussion of confidence testing cannot be complete without taking into account the argument that collecting confidence data involves more time per item than simply collecting dichotomous correct/incorrect data. Simulated confidence testing results also appear to be succumbing to ceiling effects in a way which might be alleviated by a reduction of test length (and thus an overall reduction of reliability). Further, for simplicity,

reliability will be used as the initial main outcome in order to determine where and how other outcomes might best be used.

Reliability was thus computed for subsets of the full 60 item test, decreasing in increments of 5 items. This reliability trend was computed for the traditional test as well as the 5-point and 2-point allocation methods. These results can be found in table 5, and are graphed in figure 4.

It can be seen from these results that test length does in fact have an impact on the possible benefits of confidence testing. While the traditional test takes a steady drop in reliability with the removal of items, the 5-point allocation method seems robustly indifferent to the removal of items, maintaining a reliability of .944 even if only 5 items are administered. Even the 2-point allocation only begins to drop in a substantial way when the test length drops below 20 or so items, and maintains a reliability of .802 for a 5-item test.

A different way to look at this, then, is the benefit of 2-point and 5-point allocation confidence methods relative to the traditional test over different length. This benefit by test length can be found in figure 5.

These results show that the benefits of confidence testing are most pronounced on shorter tests, and appear to asymptote to some level as the test becomes longer and longer. In fact, for the 5-point allocation line the best fit (r-square = .994) is a power curve with the specification: Y= 1.183*x^(-.691), whose limit as x-> infinity is equal to 0. This shows that in *this simulated example* the 5-point allocation confidence test will always have a higher reliability than the traditional test, and that this difference will decrease as a function of test length.

In reality it is unreasonable to believe that these results will hold to this degree on a human sample, though it does strongly support the idea that – all else equal – confidence gains will be higher on shorter tests than longer ones.

Pilot Simulation #4 – Effect of Test Difficulty

Of the other possible factors that may moderate the benefits of confidence testing, relative test difficulty is among the most clear. If a test becomes too easy for a sample of individuals they become prone to always put 100% confidence in the correct answer. Only in situations where individuals have to hedge their bets is there variance for confidence testing to show benefits.

In order to demonstrate this idea the pilot data from simulation #2 (and #3) was transformed to create two new simulated data sets. To simulate a similar but easier test, 1.5 was subtracted from each item difficulty parameter, shifting the mean from 0 to -1.5. To simulate a similar but more difficult test, 1.5 was added to each item difficulty parameter, shifting the mean from 0 to 1.5.

This was done in order to make these data as directly comparable to the original data as possible. If new item parameters were generated, differences that arose might be at least partly due to differences in the distribution of those parameters. Simply transforming the numbers to a new mean preserves their distribution.

As prior pilot data has shown, test length is an important factor in determining the benefits of confidence testing. Because of this, results were examined across the range of test length in the same way as in pilot #3. Part of these data (for a normal test) has already been shown in pilot #3 (fig. 1). Data on easy and difficult tests can be found in figure 6.

This graph illustrates a number of important ideas. First, it can be seen that the 5-point allocation method under both tests (blue lines) is more reliable than the 2-point allocation method (pink lines), which – with one exception – is more reliable than the traditional test (yellow lines). It is also the case that within each method the difficult test is less reliable than the easy test. This drop in reliability *within method* is in large part (if not wholly) due to increased guessing on the part of individuals. Increased guessing leads to a less reliable test.

This is not the whole story. The prediction of this simulation was not that the difficult test would be more reliable, but rather *that confidence testing could recover more lost information from a more difficult test.* Figure 7 shows the increase in reliability relative to a traditional test when using the 5-point allocation confidence method.

This graph shows that *the benefit of using confidence testing* is actually greatest on the difficult test. This is because confidence testing is recovering information that would otherwise be lost to guessing on the traditional test. The easy test recovers slightly less information than the difficult test relative to the normal test (from pilot #3). Interestingly enough, the benefits of confidence testing on the difficult test don't appear to drop off as quickly as the normal test with increasing test length. This implies that benefits are left to be had even on longer tests if the test is difficult enough.

Correlations with the original theta appear to also mirror these results. While the difficult test produces the lowest *r*-square values it also produces the largest gain in r-square relative to the traditional testing method. These results can be found in table 6.

Pilot Simulation #5 – Effect of Test Discrimination

Another factor that should impact confidence benefits is test discrimination. Simply put, information is lost when items are not able to discriminate adequately between those who should

get it correct and those who should get it incorrect. In these situations of low discrimination the problem is individuals of (relative) average ability who have some probability of getting the item correct, but don't clearly or consistently fall into the group of those who get items of similar difficulty correct or incorrect.

In order to demonstrate this idea the pilot data from simulation #2 (and #3) was transformed to create two new simulated data sets. To simulate a similar but less discriminating test, .20 was subtracted from each item discrimination parameter, shifting the mean from .83 to .63. To simulate a similar but more discriminating test, .20 was added to each item difficulty parameter, shifting the mean from .83 to 1.03.

This was done (as with difficulty) in order to make these data as directly comparable to the original data as possible. If new item parameters were generated, differences that arose might be at least partly due to differences in the distribution of those parameters. Simply transforming the numbers to a new mean preserves their distribution.

As prior pilot data has shown, test length is an important factor in determining the benefits of confidence testing. Because of this, results were examined across the range of test length in the same way as in pilot #3.

Part of these data (for a normal test) has already been shown in pilot #3 (fig. 1). Data on high and low discriminating tests can be found in figure 8.

Similar to the findings of pilot #4 on difficulty, the 5-point allocation method has the highest reliability, followed by the 2-point allocation method and finally the traditional test. The high discriminating tests within each method are more reliable than the low discriminating tests, as would be expected. Again, though, this is not the entire story.

The prediction of this simulation was not that the low discrimination test would be more reliable, but rather *that confidence testing could recover more lost information from it*. Figure 9 shows the increase in reliability relative to a traditional test when using the 5-point allocation confidence method.

Similar to the difficulty results, the low discrimination test has more data – relative to a normal test – which can be recovered through the use of confidence testing. The traditional high discriminating test already has a reasonable reliability, and so not as much data is being lost.

Unlike the results of difficulty the trajectory through length of test appears consistent over levels of discrimination. Thus, confidence testing benefits will decay on longer tests regardless of level of discrimination.

Correlations with the original theta appear to also mirror these results. The low discriminating test produces slightly higher benefit to *r*-square than a normal test, and the high discriminating test produces slightly lower benefit. In the case of the 5-point allocation confidence method this actually results in a doubling of the benefit (from .0562 to .1125) moving from a high discriminating test to a low discriminating test. These results can be found in table 7.

Conclusions from Simulated Pilot Data

In all, this pilot data supports each of the hypotheses proposed. Each of the confidence testing methods provided benefits to both reliability (H_P1) and validity (H_P2). As the resolution of the confidence scales increased, so did the benefits from confidence testing (H_P3). Interestingly, though, the largest gain seems to be from traditional testing to a 2-point allocation, suggesting that differences between different methods of confidence testing may not be as large as the difference between traditional testing and any confidence method. As well, this suggests

that even using a low number of allocated points may give similar benefits to much more complex methods.

Test length was a factor consistent with predictions (H_P4). This was at least in part due to ceiling effects; reliability decreased in traditional tests as items were removed while reliability of confidence test results was much more indifferent to the removal of items. Larger gains of confidence testing were found in shorter tests. This suggests that confidence tests may therefore estimate ability in a given test more quickly than through a traditional method.

Test difficulty was a factor in that more difficult tests produced larger gains from confidence testing (H_P5). This is at least in part simply due to the fact that individuals are less certain about items on more difficult tests. This leaves information available for confidence testing to capitalize on.

Test discrimination was a factor in that less discriminating tests produced larger gains from confidence testing (H_P6). Graphically this can be explained by an exercise in curve fitting – if an item is highly discriminating it will have an item characteristic curve that is very steep, and at least adequately fit by a single step function between incorrect and correct. The more that the slope of the item characteristic curve is reduced (i.e. the less discriminating the item becomes), the more steps in such a function are needed to adequately fit that curve. Traditional testing offers only one step, but in high discriminating items that is all that is needed. There is no information remaining for confidence testing to utilize. Confidence testing offers step functions with multiple steps (as many as are found at the meeting point of desire and practicality), and is therefore able to capture all the information lost by a single step function in less discriminating items.

These pilot results show some of the possible ways that confidence testing may have met its downfall, and the boundary conditions under which it might be the most useful. Unfortunately, very few confidence testing studies reported even a fraction of the information required to retroactively test any hypothesis with more modern techniques such as meta-analysis. It is no surprise, then, that none ever treated this information in experimental design.

Experimental Hypotheses

Similar to that of the simulated pilot, the main hypotheses for human subjects relate to the benefits that are associated with confidence testing. Specifically:

H1: Confidence testing will produce a more reliable test than an equivalent traditional test.

H2: Confidence testing will produce a more valid test than an equivalent traditional test.

Those findings relating to test length, difficulty, and discrimination should also carry over to human subjects from the pilot data:

H3: Benefits of confidence testing will be related to test length in that shorter tests will produce larger gains than longer tests.

H4: Benefits of confidence testing will be related to test difficulty in that more difficult tests will produce larger gains than easier tests.

H5: Benefits of confidence testing will be related to test discrimination in that tests with lower discrimination will produce larger gains than tests with high discrimination.

In addition, there are a number of hypotheses that could not be tested in the pilot, and are only reasonable to propose on a human sample. These deal mostly with how individuals will interact with these tests.

H6: Confidence testing will take more time than traditional testing for a test containing the same number of items, but reliability per unit of time will still be higher in confidence testing than traditional testing.

H7: Confidence testing will take more time than traditional testing for a test containing the same number of items, but validity per unit of time will still be higher in confidence testing than traditional testing.

H8: Test-takers will understand how to take a test using established confidence ratings.

H9: Test-takers will understand how to take a test using the general form of the confidence model in which many options are available to them.

H10: Test-takers will find confidence testing preferable to traditional testing.

There are also a number of hypotheses that relate to how individual differences may impact the use of confidence testing, based both on prior attempts in the literature (Hansen, 1971) as well as relationships discussed above. The individual differences to be tested relating to these ideas are general anxiety, test anxiety, risk-taking, and cautiousness (as a contrast to risk-taking).

H11: Trait generalized anxiety will lead to greater use of confidence testing.

H12: Trait test anxiety will lead to greater use of confidence testing.

H13: Trait risk-taking will lead to lesser use of confidence testing.

H14: Trait cautiousness will lead to greater use of confidence testing.

Finally, pilot results have shown that item difficulty plays a role in the use and effect of confidence testing. What is not to be forgotten, however, is that item difficulty alone is unimportant unless held in reference to the ability level of the sample to be tested. As discussed, it is the relative difficulty which is actually driving confidence results. Unlike in simulation,

individual test-takers do not have perfect self-concept of their ability level. Instead, perceptual information regarding an individual's understanding of their ability may in fact be important. Rather, it is not whether or not an individual is highly capable of answering the questions, but rather if they believe they are highly capable. In this way, self-efficacy may have a relationship with confidence testing results.

H15: Test specific self-efficacy will lead to lesser use of confidence testing.

Pilot # 6 – Obtaining Items and Item Characteristics

In order to replicate the above simulated results on human subjects, items need to be selected that differ both on difficulty and discrimination. These items must also be a common multiple-choice format, and understandable to the target population (college students). Unidimensionality of the items is also important in order to avoid confounds of type of knowledge.

To meet these requirements, verbal analogy items similar in format to those found on older versions of the SAT and GRE were chosen. In all, 120 items were collected from publically available SAT and GRE practice tests on the Internet. These 120 items were piloted on two groups of individuals, both from the psychology research pool at Michigan State University.

The first group of individuals consisted of 213 college student participants, and the second group consisted of 195 college student participants. No demographic information was collected from these participants, and as such we have no information on age, race, gender, or anything similar. However, given the population from which these individuals were drawn is identical to the population of the final experimental sample, it is not unreasonable to suggest that

this sample is predominantly female (~70-80%), predominantly Caucasian, and mostly between the ages of 18 and 22.

The first group of individuals received 100 analogy items. The second group of individuals received 20 analogy items (this group also received antonym items in order to test these items as an alternative to analogies; analogies performed better). Three of the items in the second collection were duplicates of items in the first collection in order to compare the relative ability of the two groups. The differences on these items were fairly small, and considered enough to assume these two samples comparable on ability. In this way they will now be treated as one sample. A more accurate difficulty and discrimination estimate for all items chosen can be recreated from the control condition of the final experiment.

Item difficulty was computed as the proportion of individuals who answered each item correctly. Item discrimination was computed as the correlation between the responses on a given item and the total scores for the test.

Item difficulty ranged from .135 to .892, and item discrimination ranged from -.064 to .724. Unfortunately, difficulty and discrimination were highly correlated (r = .728). This provides some small potential limitations, though it is also the case that effects of both difficulty and discrimination can be examined while controlling for the other. Further examination of a scatter plot of these data revealed this correlation as an artifact of a more curvilinear relationship biased by an excess of high difficulty (low numerical values of difficulty) items. This relationship should be corrected by selection of items for the following study.

In order to maximize the spread on difficulty and discrimination in a way which best allowed for the control of confounds, four groups of items were selected: 1) low difficulty, low discrimination, 2) high difficulty, high discrimination, 3) low difficulty, high discrimination, and

4) high difficulty, low discrimination. To do this, items were first ranked on difficulty and discrimination. To find items in the first two groups (low/low, high/high), their ranks on difficulty and discrimination were summed. Helped in part by the correlation between difficulty and discrimination, 15 items with the highest sums and 15 items with the lowest sums were selected. These items were also checked to ensure that they were not simply exceptionally high on one dimension but low on the other. Further, these groupings are simply to pull items that may work the best out of this sample of items. After actual data collection item difficulty and discrimination will be recalculated from the control group.

In order to select items for the remaining groups (low/high, high/low), a similar method was used. Instead of a sum, a difference was taken. Items with the highest magnitude positive (15 items) and negative (15 items) difference were thereby selected, bringing the total test up to 60 items.

While full test statistics cannot be given at this point due to the separate samples, test characteristics can be approximated using the items which were drawn from the first sample (49 of the 60 items). These items show an average difficulty (.522) and discrimination (.416), and a high reliability ($\alpha = .902$). These 60 items are therefore considered appropriate for use in evaluating confidence testing, and can be found in appendix A, along with corresponding difficulty and discrimination.

Experimental Method

Participants

Participants were drawn from the psychology subject pool at Michigan State University. Overall, 252 individuals completed the online questionnaires. Of these 252 individuals, 197 signed up for and attended the follow-up laboratory session in which test data was collected. Of

these 197 individuals, 8 were removed for bad or incomplete data relating to computer and network problems, leaving 189 participants in the full laboratory sample. This sample was predominately female (76%), Caucasian (78%), and between the ages of 18 and 22 (96%).

Upon arrival in the lab, participants were randomly assigned to one of three conditions: traditional testing, 5-point confidence testing, and a more generalized confidence testing (to be discussed more later in the paper). After removal of bad data, each condition had sample size as follows: 1) traditional testing, 65 participants, 2) 5-point confidence testing, 61 participants, 3) generalized confidence testing, 63 participants.

Measures

Analogy test.

This test can be found in appendix A, and is the direct result of pilot #6. Initial estimates showed that these 60 items possessed overall average difficulty and discrimination (.512 and .416, respectively), and good reliability ($\alpha = .902$). In addition, items were selected in order to create maximal spread on difficulty and discrimination, while also allowing for the ability to control for their confounding effects on each other.

Difficulty and discrimination as calculated from the control condition (difficulty = .56, discrimination = .25; N = 65) show different values than pilot #6, especially in terms of discrimination. Unfortunately there is no way to determine what might be different about these two samples to cause this difference, as pilot data were fully anonymous and without any collection of individual difference measures. Reliability was also lower in this sample ($\alpha = .761$) relative to pilot #6.

The problem of an unexpected correlation between difficulty and discrimination in the pilot data ($r = .728$) also appears to have been solved by the selection of specific items with a

range of difficulty, as the correlation of difficulty and discrimination has been reduced to 0.02, and examination of the scatter plot reveals a weak but expected curvilinear relationship. This confirms that the spurious correlation in the pilot data was in fact an artifact of an overabundance of difficult items. These 60 items are much more balanced across the full range of difficulty.

Means and standard deviations of the confidence ratings are not reported, as any summary statistics relating to these items have to undergo numerous levels of aggregation. For example, the average rating on each item in the 5-point condition will always be 1, as there are 5 points to distribute across 5 items. Whether an individual rates the item as 1-1-1-1-1 or 0-0-0-0-5, the mean will be the same. The standard deviation will be constrained to be between 0 and 2.33 as well (respectively for these two examples). The average number of points placed on the correct answers is simply the scores for those items. What can be examined is the general pattern of confidence, when used.

Distribution of responses for the entirety of the 5-point confidence condition can be found in figure 10, and responses for the entirety of the generalized confidence condition can be found in figure 11. The use of the zero response in the generalized confidence method was exceptionally low (as might be expected), and so it is not included in the graph. It seems that the most often used response in the 5-point method is zero. This is not surprising, as there should be a majority of answer choices on the test which are clearly wrong to a large number of individuals. This effect is not found in the generalized condition because these responses would be instead found in the 'eliminate as incorrect' response. It may at first appear that the number of 'five' responses is low in the 5-point condition, but it should be kept in mind that only 20 percent of all responses are *actually* correct.

The generalized confidence condition produces a less clear distribution, as it appears that the modal response is in fact 5. Test-takers are also over-utilizing the low end of the scale and under-utilizing the high end of the scale. This suggests that a 10 point scale may in fact not be necessary.

Manipulation check – measure of understanding.

The manipulation check consists of a five item measure administered both immediately after training and immediately prior to the end of the experimental session. These items can be found in Appendix B. The five items place test-takers in common situations they might find themselves in on any given item, and instructs them to choose optimal answers for those situations.

Trait generalized anxiety.

The measure of generalized anxiety is taken from the International Personality Item Pool (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006), and specifically those items dealing with the facet of anxiety. This scale consists of 10 items, 5 of which are reverse worded. Respondents answered each item on a 5-point Likert-type scale ranging from strongly disagree to strongly agree. The reported Cronbach's alpha reliability for the scale is .83. Cronbach's alpha from the full sample of those who took the online measures was .87. Full text of the items can be found in appendix C.

Trait test anxiety.

The measure of test anxiety is a short form of the Test Anxiety Inventory measure created by Taylor and Deane (2002). This scale consists of 5 items. Respondents answered each item on a 5-point Likert-type scale ranging from strongly disagree to strongly agree. The reported Cronbach's alpha reliability for the original scale is .87. Cronbach's alpha from the full sample

of those who took the online measures was .898. Full text of the items can be found in appendix D.

Trait risk-taking.

The measure of risk-taking is taken from the International Personality Item Pool (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006), and specifically those items dealing with the facet of risk-taking. This scale consists of 10 items, 4 of which are reverse worded. Respondents answered each item on a 5-point Likert-type scale ranging from strongly disagree to strongly agree. The reported Cronbach's alpha reliability for the scale is .78. Cronbach's alpha from the full sample of those who took the online measures was .83. Full text of the items can be found in appendix E.

Trait cautiousness.

The measure of cautiousness is taken from the International Personality Item Pool (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006), and specifically those items dealing with the facet of cautiousness. This scale consists of 10 items, 7 of which are reverse worded. Respondents answered each item on a 5-point Likert-type scale ranging from strongly disagree to strongly agree. The reported Cronbach's alpha reliability for the scale is .76. Cronbach's alpha from the full sample of those who took the online measures was .83. Full text of the items can be found in appendix F.

Trait test specific self-efficacy.

The measure of self-efficacy is a modified version of the self-efficacy scale taken from the International Personality Item Pool (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006). This scale consists of 10 items, 4 of them which are reverse worded. Respondents answered each item on a 5-point Likert-type scale ranging from strongly disagree to

strongly agree. The reported Cronbach's alpha reliability for the original scale is .78. Cronbach's alpha from the full sample of those who took the online measures was .85. Full text of the items can be found in appendix G.

Procedure

Participants completed a number of individual difference measures online before the experimental session. In addition to the measures above, participants were also asked for their high school GPA, current college GPA, and ACT and SAT scores.

After completing the online questionnaires, participants signed up for a lab session in order to complete the testing part of this study. Upon arrival in the lab participants were randomly assigned to the traditional testing control condition, the 5-point confidence testing method, or the generalized confidence testing method.

All participants received a short training (~3 minutes) on what analogies are and some strategies relating to solving analogy questions. This part of the training was identical across all groups. In addition:

Participants in the traditional testing condition received the basic analogy test in a standard fashion. They were instructed to simply select one answer for each question the same as they would on a normal test. No mention was made of confidence testing or confidence scoring.

Participants in the confidence testing condition were given a short training (~2 minutes) on the concept of 5 point confidence testing. They were told that they had 5 points for each question which could be allocated across the answer choices in any way they desired. The proportion of points they allocated toward the correct answer would be the points they received for that item.

Participants in the generalized confidence testing condition were given a short training (~3 minutes) on the concept of generalized confidence testing. They were told that they had a number of options relating to how they could respond to any given item. These options involved eliminating answers as incorrect, selecting answers as correct, or using a weighting system from 1 to 10 to differentially weight items they believed correct with varying degrees of confidence. This method is thus a combination of methods used by Coombs (1953), Dressel & Schmid (1953), and Michael (1968).

After training, but before the analogy test, participants in the confidence scored groups completed the five item manipulation check. The purpose of this was twofold. First, these five items gave participants time to practice with confidence testing before beginning the actual collection. Second, answers on these questions were collected in order to test how well participants understood the concept of confidence testing. This same five item manipulation check was also administered at the end of the session for the second reason.

All participants were also asked a question at the end of the session relating to how desirable the test they just took was in relation to other tests they had taken in the past. Following this, participants were given a full debriefing about the purposes of the study.

Tests were scored according to assumptions put forth earlier in this paper, identically to the pilot. Score on the traditional test was the number correct score, out of 60. Score on the 5-point confidence test was the sum total of points that were placed on correct answers throughout the test, divided by 5. This scales responses in this condition so that 1 point maximum can be gained on each item, identical to the traditional test. Scores on each item thus take the values of 0/.2/.4/.6/.8/1, just as in the pilot.

Generalized confidence testing presents a more complex scoring methodology. Following from scaling arguments in the development of a general model, several scoring rules can be established, following a number of assumptions that should be met. These assumptions are: 1) if all incorrect answers are eliminated and the correct answer is chosen the participant should receive 1 point, identical to the prior two methods, 2) if a correct answer is eliminated as incorrect, the participant should receive 0 points, identical to the prior two methods, 3) if a diffident response is given between x answer choices (one of them correct), the participant should receive $1/x$ the possible points (e.g. half a point if two answers, one third of a point if three answers), 4) if a correct answer is selected among a number of weighted options, the participant should receive x/X of the possible points, where x is the weight given to the correct answer and X is the summed weight given across all answers, and 5) if an answer is generically selected as correct (without weighting) among other weighted answers, that answer's weight should be set to the average value of the other weights.

This system produces a wide array of possible point values for each item, ranging from 0 to 1 and taking on a wide range of the possible numeric values between. The reader is invited to calculate the number of distinct point values possible for each single question. The author will simply acknowledge that it is sufficiently large.

Due to the (relatively) complex nature of the scoring for this method, an alternative technique for scoring was also tested. This method will be identified as generalized confidence scoring (simplified) or similar as to distinguish it from the above scoring system. This simplified method is based on the simulated result that such large degree of resolution might not be necessary in order to achieve gains to test characteristics. Instead of the above calculations involving the numeric weighting, all responses were simply coded either as 1) eliminate as

incorrect or 2) select as correct. This second category of select as correct included all answers that received weighing values; the weighting information was effectively discarded and set to equivalent. This method was then simply scored as one of two outcomes: 1) 0 points if the correct answer was eliminated as incorrect and 2) $1/x$ points if the correct answer was selected as correct, where x was the total number of answers that were selected as correct on that question.

Analysis

Perhaps the best example of problems with prior confidence testing studies is to answer hypothesis 1 and hypothesis 2 in the same way as might be found in prior studies. In this case, reliability and validity were calculated for the full 60 item test in each of the four conditions. Results can be found in table 8.

The generalized confidence test produced the most reliable results, and the simplified version of generalized confidence scoring retained almost all of the benefit of the more complex version. Using confidence intervals from bootstrapping these reliability results (1000 iterations), and the process put forth in Payton, Greenstone, & Schenker (2003), it is possible to estimate whether or not these differences in reliability are significant. At 60 items, the generalized confidence test in both forms had significantly higher reliability than the control test. The reliability on these tests was not significantly higher than the 5-point confidence test, nor was the 5-point confidence test significantly more reliable than the control.

In terms of validity, two outcomes were used: current college GPA and ACT score. SAT score was collected, but the majority of the sample (> 90%) reported that they did not take the SAT. It was therefore not used for any analysis. High school GPA was also collected, but also not used. Examination of high school GPA due to a number of unclear results led to the conclusion that a number of factors might eliminate it as a strong candidate for validity analysis.

One factor is the differential scale use due to students coming from different high schools across the country and world. Another factor is weak correlations with presumed correlates such as ACT score. This correlation is reduced to near zero when controlling for college GPA. This – and moderate correlations with college GPA – suggests that any useful variance that may be present in high school GPA may also be present in college GPA. If both GPA results are tapping the same latent ability construct, college GPA is also more proximal to this study. These factors, as well as argument for a general sense of parsimony, led to the exclusion of high school GPA from further analysis.

The 5-point confidence method performs best in relation to college GPA, though none of the coefficients are significantly different using a Fisher r-to-z transformation to perform a z test. The traditional test was the strongest of the four tests in terms of relationship with ACT score. This is unsurprising, as the items on this test were modeled after items sometimes found on the ACT. This difference was again not significant. In terms of the confidence methods, the generalized methods seemed to perform slightly better than the 5-point method in relation to ACT score. Again, simplification of the generalized method's scoring seems to have little to no effect on test outcomes.

Of course, this is not the whole story with relation to reliability or validity. It is hypothesized that test length and other factors will have a large effect on reliability and validity. Test length can be examined the same way as in the pilot, by artificially shortening the test. Table 9 shows the reliability for each test as a function of test length, and figure 12 represents this data graphically. Figure 13 shows the benefits of confidence testing relative to the control condition.

These results seem to follow the general findings of the pilot simulations, if with a bit more noise. The relationship between the 5-point confidence test and the control looks the most similar to simulation results: the 5-point confidence test starts at a higher reliability, and the traditional test almost catches up with it through the addition of more items. This has the proposed effect of altering the reported benefits of confidence testing depending on which cross-sectional test length is being examined. Two studies with identical tests will show very different results if one uses 25 items (confidence testing boost to reliability ~.20) and the other uses 50 items (confidence testing boost to reliability ~ .05). Unfortunately, the Payton, Greenstone, & Schenker (2003) bootstrapping method shows no significant differences in reliability at any test length between these two methods (5-point confidence and control).

Generalized confidence testing, both in the complex and simple forms, appears to outperform the 5-point confidence testing in all situations except on a very short test (5 items). This may indicate that the 5-point method is easier for individuals to understand quickly, while test-takers are still learning the method on the first few items of the generalized confidence test. Once the test reaches 10 items both generalized confidence testing conditions appear more reliable than the 5-point confidence test. Bootstrapping shows that this difference is significant only on the test lengths of 30 items, 35 items, and 40 items.

The control and 5-point conditions appear to be converging on the reliability of the generalized confidence method, but this has yet to occur by the point of 60 items. In fact, the reliability that the traditional test reaches at 60 items is roughly the same reliability that was reached by the generalized confidence test somewhere just above 25 items, and by the 5-point confidence testing somewhere just above 40 items. Bootstrapping the reliability of the

generalized confidence conditions relative to the reliability of the control condition shows significant improvements on all test lengths except 5 items and 15 items.

While not simulated or directly predicted, it is worth examining how the length of the test affects its validity. Figures 14 and 15 show the change in validity over test length across conditions.

No consistent interaction effects of time and condition are evident in the validity data. The control test does appear to show a greater relationship with ACT for most of the test, with the generalized confidence methods converging on the same level of relationship around 25 items and surpassing the control test (temporarily) around 10 items. This effect is fairly small, and holds only for the generalized confidence test. The 5-point confidence test appears uniformly weaker than the control test in this relationship across all test lengths.

The relationship to college GPA seems to follow a roughly opposite trend. The 5-point confidence method is strongest above 25 items before reaching the level of the control test. The generalized confidence method is slightly better than control on short and long tests, but not those of medium length. Overall, it appears that length may not be as related to the validity of confidence test as it is related to reliability.

It has been suggested that any gains to reliability and validity from using a confidence test might simply be offset by the extra time that is spent taking that test. For instance, if a 30 item confidence test takes as long for a test-taker as a 60 item traditional test, then the extra information could have just as easily been collected by giving those 30 extra traditional items. In order to examine how much longer confidence tests take for participants to complete, timing information was collected on all items across all conditions. Figure 16 shows the average

amount of time spent by participants in each condition as a factor of the length of the test to that point.

It is not surprising that the traditional test takes less time for participants to complete than each of the confidence tests. What is unexpected is how similar the 5-point confidence test and the generalized confidence test are in terms of how long it takes participants to complete them. No two means for these groups are separated by more than 15 seconds at any given test length. This suggests that the thought process that individuals are using to take the 5-point confidence test is the same or very similar to that used to take the generalized confidence test.

On average, it took participants a little over 14 minutes (858 seconds) to complete the 60 item traditional test. In this same space of time, participants were able to complete a little over 40 items in either the 5-point confidence condition (average time for 40 items = 810 seconds) or the generalized confidence condition (average time for 40 items = 814 seconds). Looking back at table 9, we can see that the reliability of the traditional test at 60 items is 0.761. This is actually quite close to the 5-point confidence test at 40 items ($\alpha = .747$) but still lacking against the generalized confidence test at 40 items ($\alpha = 0.848$).

To match the reliability of the 60 item traditional test (average time = 858), participants only needed to take 25 items of the generalized confidence test (average time = 513 seconds). This produces almost identical reliabilities (.761 vs .759), and would give test-takers on the generalized confidence test an extra five minutes to do with what they please. Even considering the additional time spent *training participants on how to take the test*, participants in the generalized confidence testing condition would still be left with an extra two minutes.

In terms of validity, it can also be suggested that when different validity scores are compared across tests, they should not only be compared by the number of items, but the time it

takes to complete any given number of items. A number of discrete time intervals can be examined, 200 seconds, 400 seconds, 600 seconds, and 800 seconds. With approximations (conservatively selected to always minimally favor traditional testing), participants can complete roughly 15, 30, 45, and 60 items in those periods of time, respectively, in the control condition. In those same increments of time, participants can complete 10, 20, 30, and 40 items, respectively, in the confidence conditions. College GPA will be examined first, followed by ACT.

Examining figure 14, it can be seen that at 200 seconds, the control test has a correlation with GPA of about .25, just about identical with the 5-point confidence test. However, the generalized conditions are both quite a bit better, with coefficients around .35. At 400 seconds this seems to shift, as the control test validity is still only slightly higher than .25, but the 5-point method has climbed to nearly .35. The generalized confidence conditions appear slightly weaker than 200 seconds prior, as they are now closer to .33. At 600 seconds, the control test begins to lose validity, and now has a coefficient of around .24. The 5-point confidence test is still quite a bit better at this time interval, as it has climbed to .37, while the generalized methods have dropped, just above .25 at this point, effectively comparable with the control condition. Finally, at 800 seconds, the control condition remains near the least valid of any of the tests, right around .24. The generalized confidence conditions are beginning another climb in validity, and are nearly back to a coefficient of .30. The 5-point confidence test is still quite a bit ahead, at nearly .40. In fact, the 5-point confidence test would actually benefit from stopping test-takers slightly earlier (at 35 items instead of 40), as validity at that point is the highest of any test at any time interval, just above .40.

This is the sweet spot, so to speak, for the 5-point method, while the optimal points for the generalized and control conditions are 10 items and 20 items, respectively (validities of .35 and .30). Neither is as high as this maximal 5-point value, which is obtained after about 700 seconds. Time spent after this point seemingly only hurts validity. Based on the general trend, no finite (or infinite!) number of similar items introduced into the generalized or control conditions would ever reach this value.

In relation to ACT, at 200 seconds the generalized methods seem to be performing best with validity coefficients of around .65. The control test at this point has a validity of around .60, and the five-point test has a validity coefficient just below .45. This, coincidently, appears to be the ideal time period for the generalized conditions, as spending more time on the test only hurts validity in relation to ACT after this point. At 400 seconds, the control test is only slightly higher than the generalized test, with coefficients of .65 and around .61, respectively. The 5-point method at this point is slowly climbing and at .50. At 600 seconds, the control test has reached a somewhat stable plateau around .70. This is higher than .60 for the generalized confidence method and .55 for the 5-point method. Finally, at 800 seconds, the control test is still at around .70, where extrapolation suggests it will remain with any additional items. The generalized test is around .65 (no different than back at 200 seconds, suggesting that those additional 600 seconds were unnecessary), and the 5-point method has also reached a plateau of just below .60, where it also appears it will stay regardless of any additional items.

In all, then, when scaling for time it appears that the generalized confidence test is ideal in terms of validity relating to ACT, as it reaches a validity of .65 with only 10 items (just over 200 seconds. While the control test eventually gets higher than this, it doesn't reach this point until around 30 items, which is a full twice as long, temporally (400 seconds).

Aside from time, there were a number of other factors predicted to influence reliability and validity. Item difficulty and item discrimination can be examined by breaking apart the overall test into subsets of items which were used to produce a test with variance on both of these factors. Due to the differences in difficulty and discrimination between this sample and the pilot sample, difficulty and discrimination were recalculated from the control condition. While the overall characteristics of the test seem to have shifted to some degree, there did not appear to be any items which changed drastically in relation to other items.

As a brief aside before moving on to these hypotheses, the computation of difficulty and discrimination from the control condition also raises an interesting (but un-hypothesized) question of the ability to calculate difficulty and discrimination from confidence scored tests.

Conceptually, difficulty is the proportion of individuals who answered correctly on any given item. Mathematically, it is simply the mean of all responses (scored 0 and 1) on that item. It follows, then, that the mean of all responses on the confidence forms (which are bounded by 0 and 1) should yield similar values. The confidence tests' difficulty scores for each item correlate very highly with the difficulty scores obtained from the traditional test (control with 5-point, r = .966; control with generalized, r = .957; control with generalized simplified, r = .957). This suggests that difficulty values generated from confidence scored tests may be able to be directly equated with difficulty values from traditional tests.

In the same way, discrimination is nothing more than the correlation between participants' scores on an item and their scores on the test. Unlike difficulty, the discrimination scores for the confidence tests do not correlate nearly as high with those from the traditional test (control with 5-point, r = .524; control with generalized, r = .358; control with generalized simplified, r = .342). This suggests that different items might be more or less discriminating

across these different score methods. Closer examination shows that on average, confidence items appear to be discriminating better than items in the traditional test (control = .252, 5-point = .266, generalized = .306, simplified = .308). It is expected that difficulty should have some impact in this relationship, as easy and difficult items should be less discriminating on a confidence test. Figure 17 shows that this is half correct.

This figure shows the trend of item discrimination relative to item difficulty, by test. Only control and generalized confidence were compared, as the effect in the 5-point condition was similar but weaker. This figure gives the clearest picture of what seems to be occurring. On the 30 easiest items, generalized confidence testing is getting more discriminating power out of each item, in some cases to a fairly substantial degree.

As difficulty increases, things become less ordered. The confidence items in general appear to be trending downward, and their relationship with the traditional items appears less clear, weakening the correlation between the two. Again, confidence tests appear to have higher discrimination overall, but this result would be drastically different if the test consisted of only easy or only difficult items. This then partly shows how prior research in confidence testing could find drastically different results from one study to the next. Difficulty seems to be having powerful effects that require more examination and potentially future work. The next set of analysis focuses on this question of difficulty and discrimination more closely, so closer scrutiny will not be undertaken yet. It may be required, however, depending on the outcomes of some of the tests of hypotheses relating to difficulty and discrimination.

Breaking apart the test into those items that are easier and more difficult as well as more or less discriminating allows for examination of the main effects of difficulty and discrimination, as well as the interactions. Figure 18 shows the difference in reliability between two 30-item

subtests, one with easier items and the other with more difficult items. Figure 19 shows the difference in reliability between two 30-item subtests, one with less discriminating items and the other with more discriminating items.

While discrimination seems to be acting as predicted in relation to reliability, difficulty is actually showing effects opposite those found in simulations. Confidence testing appears to be showing benefits on the easier items, and detriments on the more difficult items. To examine the possibility of an interaction effect, figures 20 and 21 show the effect of difficulty, discrimination, and test type on reliability.

These figures do suggest that difficulty and discrimination are interacting to some degree with test type and each other in terms of their effects on reliability, though they do not give any insight to explain the main effect of difficulty. On difficult items, all test types are weakened by low discriminating items. On easier items, confidence tests seem to be more resilient to drops in item discrimination. In fact, the generalized methods seem almost completely indifferent to this change on the easier items, with a reliability on the 15 easy/low discriminating items that is still higher than the reliability of any comparable set of 15 control items.

Given the drastically different results of difficulty in relation to the pilot, further examination of the test and of responses was undertaken. The argument from the pilot involving item difficulty is based on the assumption that easy items create a floor effect at which point all individuals should be responding with 100% confidence, completely eliminating the need for confidence scores. From there, as items get more difficult, test-takers should use confidence ratings more and more as they become less certain in their answers. In fact, data suggests at least part of this aspect of difficulty is true.

For the 30 easiest items, participants on average answered just above 8 items with confidence ratings in the 5-point condition and just below 14 items in the generalized condition. For the 30 hardest items, participants on average answered just below 23 items with confidence ratings in the 5-point condition and just below 24 items in the generalized condition. Participants are using confidence ratings more, as predicted, but the benefits are not emerging. This increased use on these difficult items actually seems to be hurting reliability.

Observed data is therefore not matching predictions of the simulations in relation to how test-takers use confidence testing on difficult items. One of the assumptions on which the simulations were built was the idea that individuals would reach a point where they had almost no hope of answering the question correctly; in this situation they would spread their confidence evenly across all answers. To further examine this, two questions were chosen at random to represent a medium difficulty item and a very difficult item.

These two items do show again that individuals are using some spread of confidence ratings more on difficult items. On the medium item only three test-takers in the 5-point confidence condition and two in the generalized confidence condition spread their confidence completely evenly. A number more, as expected, spread their confidence across 3 or 4 answers, but were still able to eliminate some choices.

On the more difficult item (in fact one of the most difficult on the test), participants should have been much more likely to spread their confidence evenly. While they did this more frequently (16 in the 5-point confidence condition, 12 in the generalized confidence condition), this is well below what would have been predicted by simulation. The difficulty on this item in the control condition was .077, meaning that less than 1 in ten test-takers were able to get this item correct. The difficulty score in the confidence collections was somewhat higher, at .138 in

the 5-point confidence condition and .134 in the generalized confidence condition. This sheds light on a number of outcomes that may be occurring on difficult items.

A difficulty of .077 means that this item was below a pure chance-based guessing parameter, which on this test would have been .20 (1 in 5). Participants in the control were doing worse than blind guessing, indicating that a good distracter option on this item was likely present. Distracter options were not examined in simulation; at a certain level of difficulty participants were simply assumed to be answering with full randomness. This randomness manifested itself as error in simulations of traditional tests, and this error was tempered by confidence testing, resulting in gains. In simulation, there were no differential assumptions made about items when difficulty dropped below 1/*k*, and this neglects the fact that in a human sample a break occurs at the point in which individuals begin responding worse than chance. In fact, if more test-takers were using confidence testing rationally, and did not see or fall for distracter items, the difficulty of all items below a certain threshold should bottom out at 1/*k* (or .20 in this case). This tendency of the scores to cluster at this level is actually evident in the slight rise in computed difficulty in the two confidence collections relative to control.

Mathematically this is expected. Individuals who are spreading their confidence in a way similar to traditional tests would be producing an item with difficulty near .07, and those using confidence testing to spread their confidence evenly should be producing an item with difficulty near .20. It is thus not surprising that the computed difficulty of this mix falls somewhere in between, near .13.

Overall, this suggests that the effect of difficulty found in simulation is valid only on tests where good distracter choices do not exist, or more specifically, on tests where item difficulties do not naturally drop below 1/*k*. This can be empirically examined, within limits, as 8 of the 60

items have difficulties that fall below .20. Reanalysis on 52 items after removal of these 8 does show *minimal* improvement in getting confidence tests back to zero loss on difficult items, but not enough of an effect to dwell on much longer. It should go without saying that it is far from reversing the effect to line up with simulation results.

Part of the problem in testing this specific possibility is something that has been discussed many times to this point in this paper. Item difficulty should not be only considered as a fixed property in a vacuum, but rather a level that each test-taker tries to surpass with their individual abilities and differences. An item with difficulty of .22 still shares the same problems as one at .18, just slightly lessened. There is a gray area in which the distribution of relative difficulty drops below this difficulty threshold for some proportion of the sample. For the sake of brevity, then, let us cut to the chase and examine this effect on items which should not have this problem.

The 20 easiest items on the test (difficulty ranging from .831 to .969) were divided into two subtests of 10 items each: the easiest items and the slightly more difficult items. Reliability on these sets of items does actually look closer to simulations, and can be found in figure 22.

While this does follow much closer to simulation, and does suggest that confidence testing might have its greatest benefits on items that are not too difficult or too easy, these results must be taken with some healthy skepticism. This is a small set of items relative to the full test, and confidence testing has already shown to have benefits associated with test length. Test length is accounted for in this examination, but it is still a small number of items. What this does seem to show is that the predicted ceiling effect of easy items seems to be a minor factor for the generalized confidence condition and a slightly larger factor for the 5-point condition.

The impacts of difficulty, discrimination, and test type on validity produce twice as many results as reliability due to the fact that two different outcomes were validated against. Overall, results do not appear to support predictions. Differences appear to follow somewhat consistently with those results reported for reliability (but in general, weaker), with confidence tests performing better on easier items and those with higher discrimination. The effect of discrimination appears to be much weaker, with the traditional test not taking as large of a loss on low discrimination items. High difficulty, low discrimination items still appear to show the weakest validity for all test types. Due to the lack of firm results, the similarity with reliability, and for the sake of simplicity and brevity, results will not be explicitly discussed (though are available upon request from the author).

Hypothesis 8 and 9 predicted that participants would understand how confidence scoring works and how to use it on the items in the manipulation check. Due to the different nature of 5-point confidence testing and generalized confidence testing, each will be examined in turn.

The manipulation check on the 5-point confidence test yields fairly quantifiable correct and incorrect answers. For example, participants should respond with '1' to all answers on the item which asks them to respond as if they had no idea which answer was correct. Overall, there were very few deviations from these 'correct' answers.

The first item asked participants to 'Answer this question as if you couldn't decide between answer choice 'A' and answer choice 'B', but felt slightly more confident in answer choice 'B'.' Of 61 participants, 91.8% in the pre-test and 98.4% in the post-test responded with the expected response of 2/3/0/0/0. Another 4.9% in the pre-test and the remaining 1.6% in the post-test responded with the close response of 1/4/0/0/0.

The second item asked participants to 'Answer this question as if you were fairly certain that answer choice 'C' was correct.' Of 61 participants, 86.9% in the pre-test and 91.8% in the post-test responded with the expected response of 0/0/5/0/0. Of the remaining, the most common mistake was x/x/4/x/x, where a 1 was substituted for one of the 'x's.

The third item asked participants to 'Answer this question as if you had no idea as to which answer choice is correct.' Of 61 participants, 98.4% in both the pre-test and the post-test responded with the expected response of 1/1/1/1/1.

The fourth item asked participants to 'Answer this question as if you were certain that answer choice 'D' and answer choice 'E' were incorrect, but had no strong idea of which of the remaining was correct.' Of 61 participants, 85.2% in both the pre-test and post-test responded with the expected response of x/x/x/0/0, where 5 points were distributed as equally as possible (i.e. 2/2/1, 2/1/2, 1/2/2) across the answers marked with x.

The fifth item asked participants to 'Answer this question as if you were fairly certain answer choice 'A' was correct, but had some thoughts that there was a chance that answer choice 'B' was correct.' Of the 61 participants, only 65.6% in the pre-test and 73.8% in the post-test responded with the expected answer of 4/1/0/0/0, though an additional 31.1% in the pre-test and 23.0% in the post-test answered with the close response of 3/2/0/0/0.

The manipulation check on the generalized confidence test has a wider range of possible answer strings that would work for each question, though each can still be quantified as fitting with the instructions of the question or not.

For the first question, 60.3% in pre-test and 74.6% in post-test responded in an appropriate way. For the second question, 68.3% in pre-test and 63.5% in post-test responded in an appropriate way. For the third question, 82.5% in pre-test and 81.0% in post-test responded in

an appropriate way. For the fourth question, 85.7% in pre-test and 92.1% in post-test responded in an appropriate way. For the fifth question, 74.6% in pre-test and 73.0% in post-test responded in an appropriate way.

Overall, these results are not as promising as the 5-point confidence test. However, the majority of errors on these items seem to be cases where participants left some of the answers blank on a given question. Fortunately, and for whatever reason, this does not appear to be as much of a problem on the main test as on these manipulation check questions. Practically, this may inform that a choice such as 'eliminate as incorrect' is not needed, as the answers left blank often appear to be those that should have been eliminated. This is a possible way of simplifying administration of this test method and helping to bring about better understanding of it.

Hypothesis 10 predicted that participants would find confidence testing preferable to traditional testing. All participants were asked a question asking them to judge the test they just took in terms of desirability relative to other similar tests that they have taken. Responses were given on a 5-point scale from 'Much less desirable' to 'Much more desirable'. One-way ANOVA shows group differences on this question ($F(2,188)=11.903$, $p<.001$), with least square difference post-hoc tests showing that participants found the 5-point confidence test (M= 3.59, SD= .761) more desirable than either the traditional test (M= 2.89, SD= .732) or the generalized confidence test (M= 3.16, SD= .919). While participants did not find the generalized confidence test significantly more desirable than a normal test, they at least found it no more objectionable.

Due to the arguably quasi-continuous nature of test desirability as measured, a chi-square test of association was also performed examining the relationship between test and desirability. This test was also significant (χ^2 (8,189) = 26.62, $p<.01$) and supportive of the results of the ANOVA.

It was predicted that a number of individual difference measures would influence how individuals interacted with confidence tests. Specifically, these individual differences were predicted to change the degree to which test-takers would use confidence testing to distribute their confidence across answers, instead of simply placing all their confidence on one answer. A new variable was created in each of the two experimental datasets which represented a count of the number of items on which each participant used confidence testing to distribute points across multiple answers. Table 10 shows the correlation of each of the individual difference measures with the number of items on which confidence testing was used for the 5-point confidence test, and Table 11 shows the same for the generalized confidence test.

Results are somewhat mixed, with no significant correlations but some results trending in the predicted directions. The generalized confidence condition shows a marginal effect of test anxiety in that participants higher on test anxiety are more likely to use confidence testing (as predicted).

There is, however, a potential problem with measuring the 'use of confidence testing' in this way. That problem relates to ideas discussed in the simulations of the effect of difficulty on the benefits of confidence testing. It was suggested that confidence testing may not show benefits on easy tests due to the fact that on very easy tests participants would always (correctly) display 100% confidence in the correct answers. As difficulty is relative to the person, it stands that the number of times that a person uses confidence testing should be related to the difficulty of the test, or rather, how difficult the test is relative to their level of ability. This correlation between score on the 60 item test and the number of times that a person used confidence testing is significant in both the 5-point confidence condition ($r = -0.573$, $p < .001$), and the generalized confidence condition ($r = -0.595$, $p < .001$). This is an expected result and not relevant to the

predictions of the individual difference measures. It should therefore be controlled before examining the effects of the individual difference measures. Partial correlations identical to the above analysis but controlling for participants' score on the test can be found in tables 10 and 11.

Overall, all effects appear to be negated when controlling for a participants' score on the test. To examine the possibility that some individual differences might be acting the same way across the forms, and to increase the power of these correlations, the same partial correlations were examined on a dataset containing responses from both the 5-point and generalized confidence condition. These partial correlations can be found in table 14.

Even with this increased sample size, there do not appear to be any relationships between individual differences and the use of confidence testing in general.

Discussion

The following discussion of results will be grouped to first focus on reliability, then validity, then individual differences and person factors. In general, results will be summarized first, then examined more in detail relating to practical implications, limitations, and areas for future study.

Hypothesis 1 proposed that confidence testing would produce a more reliable test than a traditional test. This hypothesis is generally supported, with both the 5-point confidence and generalized confidence test producing higher reliabilities than traditional tests of the same length. However, this difference in reliability (while similar in magnitude to prior results), is only significantly higher in the generalized confidence condition. Also noteworthy is that a simplification of the scoring methodology for the generalized confidence test appears to have negligible impact on reliability in relation to the full generalized confidence method, showing that a simpler system using this same general form may be ideal.

Hypothesis 3 proposed that the benefits of confidence testing on reliability would be related to the length of the test in that shorter tests would show larger benefits than longer tests. Discounting odd effects at extremely short test length (~5 items), it does appear that test length impacts the benefits of confidence testing similarly to simulated pilot data, supporting this hypothesis. As test length increases it appears that both forms of confidence testing are beginning to converge on zero benefit over traditional testing. This has not occurred by 60 items, though extrapolating by trends over number of items it is reasonable to assume this may occur around 75-80 items for the 5-point condition and 90-100 items for the generalized confidence conditions.

Hypothesis 4 proposed that confidence testing would show greater benefits on more difficult tests. This hypothesis was not supported, and in fact results appear to be opposite predictions. Confidence testing performed better than the traditional test on easier items, and worse on more difficult items. This effect was stronger in the generalized confidence conditions.

This finding may indicate a number of possibilities, but overall suggests that the relationship between test difficulty and the usefulness of confidence testing may not be as simple as simulations would indicate. It was shown that participants use confidence testing more when items become more difficult, and overall results show that confidence testing has its benefits. On the most difficult items the use of confidence testing is actually hurting reliability, which means that test-takers must be using it incorrectly or irrationally in these situations.

This seems to relate at least somewhat to distracter choices on items that lower difficulty of the item to below 1/*k*. One of the missing assumptions of the simulations was that at a certain point test-takers would reach a level of difficulty at which they would find it best to evenly distribute their confidence and cover all bases; below this point individuals would stop

displaying knowledge and instead start displaying lack thereof. In simulation, score was assigned probabilistically without taking this disconnect into account. Good distracter choices are beneficial to traditional tests, but results seem to suggest that they cause problems with individuals' consistent use of confidence testing. These difficult items decrease the number of people evenly spreading confidence. On the most difficult items the rate of this behavior is well below what should be expected, given the extreme difficulties. Even removing some of the most difficult items does not change this outcome much. The problem is that this test was not constructed with distracter choices as an experimental consideration. Instead, they are simply a confound, and one that is difficult to extract without intense analysis of item content relative to ability levels of the participants.

The differences are small, but difficult items are the one place that the simplified version of generalized confidence testing actually outperforms the generalized confidence testing as collected. This hints at one of the drivers of this problem; on the hardest items participants seem to be over-thinking things instead of simply evenly distributing their confidence. Through this process of making things more complex, they are adding noise to the data. This reduces reliability (and validity through measurement error). In the generalized condition this might result in more use of (incorrectly applied) differential weighting, which is basically 'fixed' in the simplified version. Even so, this is a small piece of the puzzle.

Some examination of less difficult items does show relationships more similar to those found in the pilot studies, but the unfortunate truth is that this study was not set up to examine these issues in detail. These results highlight paths for future research, but can do little more in terms of establishing firm conclusions of what might be found. For now, it appears that

confidence testing is weakest on extremely difficult items, notably those that are well-written enough to drive the difficulty below chance guessing.

Confidence testing does seem to work well on items of easy and medium difficulty, and the potential ceiling effects of very easy items do not appear to pose as much of a problem as predicted. This may simply be because even easy items are not easy to every test-taker. Overall, this raises the possibility that confidence testing's relationship with difficulty may be more bell-shaped than monotonic, with both very easy and very difficult items producing problems of different degrees.

Unfortunately, this test lacked groups of items at comparable difficulty, and rather filled a range of difficulty. Examining difficulty as a full spectrum thus involves examining individual items or increasingly small sets of items, something that this study was not designed to be able to do. Due to this it is hard to make comparisons about particular levels of difficulty without getting entangled in individual items, which introduces a host of other problems. Future targeted work (and ideally, an eventual meta-analysis after the collection of a number of primary studies) would be required to fully address this problem.

The findings on difficulty also raise a question of what confidence really means, especially when the items are more difficult. It is assumed that confidence is being measured through the collection of confidence data on these tests, though it is impossible to say with this data that confidence is *actually* what is being measured. Given findings, it does seem the most likely that confidence is what is driving responses, but it also possible that deviation from a pure measure of confidence on difficult items may be the driver of the decline of their usefulness. Confirmation of this would require deeper work, examining the thought processes and behavioral

motivations that lead test-takers to select one level of confidence over another. Ideally, this would likely manifest as a verbal protocol study.

Hypothesis 5 proposed that confidence testing would show greater benefits relative to control on tests with lower discrimination, and lesser benefits on tests with higher discrimination. Data supports this hypothesis. Confidence testing seems to show no benefit to reliability on the high discrimination items (all tests do well). On the low discrimination items the benefits are fairly large, especially in the generalized confidence condition. This implies that the need for highly discriminating items is not as great on a confidence test as it is on a normal test. That is not to say that items should have no discriminating power whatsoever, but rather simply that the demands may not need to be as great.

Hypothesis 6 proposed that gains in reliability are more than simply the product of adding more time to the test. Put another way, a 60-item traditional test should not be compared to a 60-item confidence test, but rather to a test that takes the same amount of time to complete. Results show that the reliability per unit time between the 5-point confidence test and the control is fairly negligible, but the difference between the generalized confidence and the control is well in the favor of the generalized confidence test.

The 5-point confidence testing method may not show any improvement over a controlled period of time, but that is not the only factor by which it should be judged. The fact that each single item (without any modification to content of the item or additional effort in its creation) can collect more information is what is also important, especially in the context of test security and item exposure. For example, allow that participants can complete 1) 60 items of a traditional test with fixed reliability in a certain time period, or 2) 40 items of a 5-point confidence test with the same reliability in the same time period. If there is no difference to reliability or time then

the only factor of note is that those in the 5-point confidence collection were only exposed to 40 items while those in the traditional test were exposed to 60. This would lower the requirements of item pools, or simply make item pools of the same size more powerful. Thus, hypothesis 6 is strongly supported in relation to the generalized confidence condition, and supported within certain assumptions in the 5-point confidence collection.

Hypothesis 2, hypothesis 3, 4, 5 and hypothesis 7 proposed effects of confidence testing on validity of the test. The test of hypothesis 2, which proposed that confidence tests would show higher validity than traditional tests, showed inconsistent results. Confidence tests showed better prediction in relation to college GPA, but weaker in relation to ACT score. The test of hypothesis 3, which proposed that test length would have an interaction effect on these validity results, also showed inconsistent results. Test length seemed to have no easily interpretable impact on the validity of the test, which also means that no fair or strong statements can be made about hypothesis 7. The simple addition of items did not appear to change validity. Changes were more likely related to the addition of specific items which themselves showed better or worse validity. Hypothesis 4 and 5, which proposed that test difficulty and discrimination would have an impact on validity mirror the results already discussed relating to reliability for difficulty, but had no notable effect in relation to discrimination.

Hypothesis 8 and 9 proposed that test takers would understand how to take a test using each of the different confidence methods. Overall this is supported in both conditions, as the majority of participants seemed able to accurately use this method of testing with even fairly minimal training and experience, though there is a question of whether this process breaks down on very difficult items. It is unclear (and an important question of future work) how many testing sessions it would take for all individuals to become acclimated to using these confidence

ratings with complete accuracy. Even minimal one-on-one training would likely be very useful to those who did not immediately understand the process through group training.

Hypothesis 10 proposed that individuals would find confidence testing preferable to traditional testing. Contrary to many of the other score-based results which showed generalized confidence testing outperforming the 5-point method, participants found the 5-point confidence condition to be the most preferable. There was no significant difference between the preference toward generalized confidence testing and traditional testing. This provides partial support for hypothesis 10, though an argument can be made that if generalized confidence testing is at least not *less* preferable than a traditional test, its benefits come at no psychological harm to the test-taker.

The remaining hypotheses related to how individual differences influence the use of confidence testing. Hypothesis 11, 12, 13, 14, and 15 all failed to be supported by the data. Neither confidence condition, nor the combination of the two, showed any significant relationship between the use of confidence testing and scores on individual difference measures. This failure to find any result is not quite as disheartening as it may at first appear. In fact, for confidence testing to be a more widely utilizable method it is preferable that its use is not based on outside factors. Risk-taking, which was found by Hansen (1971) to be somewhat correlated with use of confidence ratings ($r = -.226$, $p < .05$), appears to have a weaker effect in this sample. A larger sample size may have found some of these smaller effects to be significant (note: the overall sample size of the two confidence groups was still larger than Hansen's 1971 study), but it is at least possible to rule out any medium to large effects of these variables on confidence testing use. If effects do exist, they are sufficiently small as to not have been picked up by this study, or represent other constructs that weren't measured.

It may also be the case that different situations may prompt more individual difference effects. For example, it may be that context is important in measuring some of the individual differences, such as self efficacy. Instead of test-specific self efficacy, *verbal* test-specific self efficacy could be examined in future work. These effects may also be weaker due to the fact that this test was low stakes. On a high stakes test these effects may become larger and significant, as there may be psychological processes that were not activated by this simple low stakes setting. While individuals were told to imagine that they were taking a normal test as they would in a class, there is no substitution for actual high-stakes data. This examination of individual differences would also benefit from more in-depth examination of confidence as it stands itself while individuals are taking the test, as might be gained from a verbal protocol analysis.

It is also possible that other constructs that were not examined may have some bearing on the use of confidence testing. These constructs were examined as a first pass at finding covariates, but they hardly cover the full range of possible individual differences that might be related to testing. The case might also be made that these constructs were simply measured poorly, though such speculation seems unlikely considering the relatively high reliabilities produced by each of the measures.

Practical Implications

Perhaps the main question that needs to be answered by this paper is: "should confidence testing be implemented in practical settings?" The general answer appears to be yes. This of course comes with some call for caution and continued scrutiny, which will be discussed relating to necessary future research, but confidence testing does stand to provide some practical gains to testing.

An additional question seems to also be "which version of confidence testing is best, or at least most practically useful?" Each test seems to excel in different areas, and while the benefits of generalized confidence testing appear larger, so are the detriments when things don't work out (e.g. highly difficult items). In general, it is the opinion of the author that the best hope is likely to be found in the simplified generalized confidence method, but collected in a simple form instead of with the 10 point weighting. Even if this collection only had the same benefits as the simplified form in this study, test-takers could likely complete it quicker, increasing the gains when temporally scaled. Additionally, this simpler form might give test-takers less ability to act irrationally on difficult items. Unfortunately, this can only be examined with future work.

First and foremost, college students with minimal training were able to figure out and use these techniques with a large degree of accuracy. Training can be implemented using a PowerPoint presentation in a matter of minutes, which allows for fairly easy implementation if that is what is required or desired. Additional training could be used to improve understanding, which would also likely grow with more widespread use. Anecdotally, only one participant from both confidence conditions (it was in the 5-point confidence condition) asked the experimenter for clarification of the technique, and even then only on the first of the practice questions. Participants didn't dislike the methods any more than a normal test, and individual differences did not seem to have any moderate or large impacts on usage of the test. The issue that confidence testing takes more time than traditional testing seems somewhat unfounded, as controlling for time did not cause any substantial detriments to the techniques that would make them worse than traditional testing.

Above all this, confidence testing does seem to show some noteworthy improvement to tests. The 5-point confidence method shows some boost to test reliability, and even though this

benefit is reduced to near-zero when accounting for the time spent on test it still provides a practical use in terms of test and item security as outlined above. The generalized confidence method showed a boost in reliability even accounting for time spent on item.

This study also provides a worst-case scenario in terms of time spent on each item, as it is a fair assumption that this is the first time these test-takers had ever encountered confidence testing methods. With practice it would be expected that test-takers should get better at taking tests in this fashion, perhaps strengthening time effects in favor of confidence methods.

Validity results were slightly less clear than those of reliability, but still offered some hope for the techniques. Both methods of confidence testing show signs of offering better validity than traditional testing in relation to college GPA, but both appear slightly weaker than traditional testing in relation to ACT score. There are two possible explanations.

The first is that the ACT can be considered a benchmark measurement of verbal reasoning. The fact that confidence testing does not correlate as highly with this as the traditional test (which is effectively a replication of the ACT), would then mean that the confidence method is introducing more noise which is not important. Oddly, though, this extra noise spuriously correlates better with college GPA than the replicate of the ACT (the traditionally scored test) does.

The second possibility is that the confidence testing is better predicting college GPA because it is capturing more useful variance in college GPA, In fact, it is capturing more variance in college GPA than the traditional test, which is again basically a replication of the ACT. By correlating better with extra variance in college GPA that ACT isn't measuring, the correlation with ACT is lowered. This information that confidence testing is capturing is only viewed as noise in the correlation with ACT in that the ACT fails to capture this useful variance.

Further, given that ACT score *is itself* a tool meant to predict college GPA (for reference the correlation between the two in the full data is $r = .420$, $p < .001$), it may be that the prediction of college GPA holds more weight in this study. These initial results are therefore somewhat promising, but call for continued work. Future studies can potentially utilize tests which are more directly related to outcomes. Verbal reasoning is a fairly large construct, and one that has the potential to correlate with a number of factors which may not be verbal reasoning. More specific tests of ability (such as task-specific performance or specific tests of skill) may allow a cleaner examination of validity results.

Hypotheses relating to difficulty did not work out exactly as predicted by simulations. From a practical standpoint, this means that use of confidence testing should continue to use a wide array of items of varying difficulty as a means of further understanding these impacts (and protecting against weaknesses of a test using limited ranges of these characteristics). There are no singular set of item types that demonstrate cause to be selected over all others, though it appears that items developed to have good distracter choices might be best left out of non-research use. While it seemed that optimal difficulty for confidence testing was near .70 in this study, this finding should be replicated in other samples with other tests. Additionally, as difficulty is relative to a number of factors it may take a number of targeted studies to determine exactly what relationship item difficulty has with the benefits of confidence testing, and how exactly test-taker behavior breaks down at the difficult end of the scale. The addition of some difficult items to any collection may be worth the extra effort for purposes of study even though they appeared less than ideal on this test and sample.

Deeper perceptual studies examining what confidence means to test-takers across levels of difficulty may also shed light on this issue. As discussed earlier, it might be the case that

confidence is being measured by these confidence items only on easy and medium difficulty items, but not on the most difficult. It may also be that confidence is not being measured at all. There appear few ways to address this other than in depth verbal protocol analysis, and this may be required to fully determine what exactly is being measured by these ratings from a perceptual standpoint.

Hypotheses relating to discrimination were supported in relation to reliability, and tend to indicate that confidence tests (and especially the generalized confidence form) are somewhat indifferent to item discrimination. The generalized confidence test used here returned only a slightly weakened reliability on the 30 least discriminating items compared to the 30 most discriminating. The traditional test was reduced from a good test on the most discriminating items to completely unusable on the least discriminating items. This then reduces the ROI of time and effort spent making sure items are maximally discriminating. The author is by no means suggesting that items should be built without care in any aspect, but simply that less time and effort may need to be spent fine tuning decent items to make them into good or great items. On a confidence test those same items might already perform at the level of good or great.

There are also a number of practical areas in which confidence testing might show even greater use than found here. Most notable would be computerized adaptive testing. If confidence items collect more information in each item, then better estimation of a test-taker's ability level can be made earlier in the test. Mindful of stated problems with the current understanding of difficulty's relation to confidence testing, it is also the case that at least some participants would use an evenly distributed confidence weighting instead of guessing, reducing the problems of breaks in predicted Guttman response to items. As discussed above, confidence

testing would also reduce the strain on item pools, as fewer items would be needed to reach the same results. Item exposure would thus be reduced, increasing test security.

Confidence testing also creates a higher resolution pattern of data which could be useful in the detection of invalid response patterns and cheating. A Guttman error in a response string is a Guttman error – the more you have the more you may begin to suspect that a participant had prior knowledge about some of the harder items. Confidence ratings introduce another level to this. Someone with prior knowledge of the test would not only be getting the most difficult items correct (which could be a product of good guessing on a traditional test), but may also be getting them correct with full confidence – a much less likely prospect. A cheater would then face the dilemma of having to report less than full confidence (and receive less than full points) on some items in order to keep their test from increased scrutiny.

As a teaching tool, confidence testing may have potential uses even without reliability or validity benefits. Looked at as a shortcoming in terms of test characteristics, let us dwell on the notion that *test-takers spent more time taking the confidence tests than the traditional test*. Participants weren't simply sitting there and wasting time, nor is the extra time they took fully explainable with only the extra mouse clicks required. Participants who used confidence ratings were, in whatever way, thinking more about the items and their responses. This test was likely not perfect for it due to the lack of possibility of learning anything while taking the test, but other tests – especially in a classroom setting – could be constructed to make the most use out of this extended period of test-taker engagement. Additionally, an instructor could utilize student data to see where common mistakes or areas of uncertainty might be occurring in order to inform future teaching plans.

Limitations and Future Research

In addition to a number of positive findings, this study also does have a number of limitations. These limitations highlight a number of prime areas and ideas for future study. This study used a low-stakes test on a sample of college students. There is absolutely no guarantee that any of these results would transfer to a high-stakes test, and it is impossible to say to what populations these results could be generalized. These results should be tested on populations other than college students, with different types of tests, and in higher stake situations. It would be useful to know if these forms of confidence tests worked on younger students or on average members of the workforce, which would allow for implementation in schools, and selection and workplace situations, respectively. Testing in actual classroom studies would also allow for examination of how test-takers use these methods in higher stake situations. High stakes tests might also change the impacts of individual difference measures on confidence ratings.

This study used also used a test that was narrowly focused on the idea of verbal reasoning in the form of analogies. It is unclear if these same findings should be expected for a reading or math test, and the different ways that individuals view and approach verbal and math tests might suggest that differences may appear. Replication on other types of tests would be useful to show that confidence testing works in a variety of situations. A study could use a test similar to this one, but also comparable tests (such as math test, or a subject specific test), to allow for testing differences between these tests. Hopefully these differences are minimal, if they exist at all.

It is also unclear from this study how these results would hold or change over time as individuals became more familiar with how to take a confidence test. Future study with multiple tests over time on the same sample thus stands to answer a number of questions. Participants could be administered a series of parallel forms over the course of a few months, similar to what

might be expected in a classroom setting (or even using a classroom setting). This would allow for test-takers to become accustomed to the method of confidence testing beyond the short period of time they were involved in this study. Examination of confidence tests after multiple sessions of taking confidence tests would give a better picture of what day-to-day results of confidence testing are likely to be on informed populations. As with all tests, individuals will no doubt soon find the best strategies for using confidence testing.

Multiple tests over a span of time would also allow examination of whether or not test-takers internalize more knowledge from the test as a product of making confidence ratings. Specifically, this would be able to show if test-takers' scores change differentially as a product of taking confidence tests relative to control groups taking traditional tests. Individuals should learn information over time, but it could be tested if that degree of learning is greater in confidence testing conditions. If it is the case that confidence produces larger gains in learning that becomes a larger selling point for using confidence tests as a teaching tool.

This study used a test with 60 items, which seems to be of reasonable length for keeping test-taker attention and motivation. However, examination of aspects such as difficulty and discrimination (and their interactions) reduced the length of the tests used for comparison at the cost of being able to make more comparisons. This sacrifice of depth for breadth was necessary in this study but limits the extent to which some of these effects can be reasonably examined.

Notably, the future use (and careful documentation!) of tests of varying degrees of difficulty relative to the sample would also be a good step in untangling exactly how test difficulty is related to use and benefits of confidence testing, and what the reasonable boundary conditions of use are. A study in which groups of test-takers complete tests of fixed and more stable difficulty, but varying in difficulty between groups, all with confidence testing, would

likely produce larger effects that would be easier to detect and clearer to interpret than the spectrum of difficulty in this study.

This test was also limited by the fact that items with good distracter choices present were not experimentally controlled – this demands careful future study. With planning it would not be difficult to use good distracter choices as an experimental manipulation. Two groups could take the same test items, but good distracter choices could be implemented on one test and not the other. Care would also need to be taken to ensure that distracter choices' effect on discrimination was controlled for during this implementation. Distracter choices should have an impact on item discrimination, though careful and specific construction, manipulation, and selection items can disentangle this component to allow for a clear picture of the effect on difficulty. It is also possible that a selection of some items that do have impact on discrimination could also be examined at the same time in order to determine if any notable effects on discrimination exist. This is the cleanest way of attempting to understand these impacts, as conclusions drawn from this current study are exploratory at best.

This study is limited by the fact that it only examined two forms of confidence testing, though these were suspected to be among the most powerful. There are therefore a number of future research possibilities related to how different forms of confidence testing may work, and which may be ideal. This study was able to find two forms of data collection that worked relatively well, but there are plenty more in prior literature. These two were picked with the hope they might be among the strongest, but this by no means should limit work to try to show that other implementations are stronger. Future studies could be implemented in a similar fashion than this one, but with different methods of confidence testing.

For example, generalized confidence testing, as collected, appears to not be worth the extra effort beyond collecting a simplified version. The benefits of the generalized condition are also present in the simplified generalized condition. That said, it may be that the simplified version only worked because it was a distillation of a more complex collection. The simplified condition was just that, a simplified version of a more complex collection. If individuals did not actually have the 10 point scale to work with, it is hard to say just how they might act. This study is limited in its inability to examine this question in full – a high priority is collecting data using this simplified method to see if it works as well when collected in this form instead of simplified to this form. That is, collecting data where test-takers have the possibility to select answers as correct and eliminate answers as incorrect, but not to rate them further on any additional scale. It is possible that test-takers may find it, in the simplified form, more desirable than the full generalized confidence form. If a collection of the simplified version replicates the benefits found here it would be extremely easy to implement in almost all testing situations.

The sample size of this study, while sufficient to show most effects with some confidence, was only large enough to rule out medium and large effects of individual differences on the use of confidence testing. It is argued that smaller effects may not be large enough to cause problems, but this study does not have the power to say that they do not exist. It may also be the case that other individual differences that were not measured here have impact on confidence, so future work should include other individual difference measures when the theoretical case can be made for their inclusion. The fact that this test was low stakes might also have activated different psychological processes than would normally be activated on a higher stakes test, as would be found in a classroom setting. Replicating these individual difference results on higher stakes tests and in different settings would be useful in understanding how

individual differences may or may not impact confidence ratings. Future work should examine how individual differences (and other measures of those differences) impact the use of confidence testing, with the collection of larger sample size a high priority.

Due to the distinct aspects of a number of these limitations, a number could be studied in parallel. Some prioritization may still be helpful, however. The highest priority is likely the concrete establishment of one or two forms of confidence testing that work well. These might be the two from this study, or a collection of the simplified version of generalized confidence testing. A number of small studies could prove useful before moving on to other topics. The next highest priority is understanding just how difficulty is impacting confidence testing. This can be accomplished through a number of the above studies – the first studies should examine full tests of different difficulty, the experimental manipulation of distracter choices, and perhaps even more simulation in order to see if observed behaviors can be properly modeled. After difficulty is better understood, generalizability is a clear next priority. Showing that confidence tests can be used in a wider range of situations (or how they might function differently in different situations) would help to create more practical uses for these methods. Once these topics are better understood, final priority would likely fall to administering multiple tests over time and a larger attempt to understand how individual differences might play a roll.

It is tempting to say that this paper raises as many questions as it answers, and that may well be the case. It is argued that these new questions have at least some direction, and provide concrete steps that can be undertaken to understand confidence testing in new ways. Prior studies did little to understand why results were inconsistent. This study found a number of reasons to explain just why they might have been. Test difficulty seems to be one of the largest

factors, and future studies will hopefully shine light on exactly how this can be taken into account to provide the best possible outcomes for test writers, test takers, and test users.

APPENDICES

Appendix A – Test Items Selected From Pilot #6

Difficulty and discrimination noted in parentheses (difficulty; discrimination).

1. Bird : Nest :: (.49; .36)

   Dog : Doghouse
   Squirrel : Tree
   Beaver : Dam
   Cat : Litter Box
   Book : Library

2. Dalmation : Dog :: (.75; .29)

   Oriole : Bird
   Horse : Pony
   Shark : Great White
   Ant : Insect
   Stock : Savings

3. Maceration : Liquid :: (.05; .04)

   Sublimation : Gas
   Evaporation : Humidity
   Trail : Path
   Erosion : Weather
   Decision : Distraction

4. Bellow : Fury :: (.69; .39)

   Snicker : Hatred
   Hiss : Joy
   Giggle : Dread
   Yawn : Excitement
   Gasp : Surprise

5. Mason : Stone :: (.77; .30)

   Soldier : Weapon
   Lawyer : Law
   Blacksmith : Forge
   Teacher : Pupil
   Carpenter : Wood

6. Toe : Knee :: Finger : (.80; .41)

   Arm
   Wrist
   Elbow
   Hand
   Shoulder

7. Repel : Lure :: (.49; .27)

   Dismount : Devolve
   Abrogate : Deny
   Abridge : Shorten
   Enervate : Weaken
   Miscarry : Succeed

8. Fierce : Timid :: Aggressive : (.85; .35)

   Weird
   Frigid
   Assertive
   Cold
   Passive

9. Coax : Blandishments :: (.48; .16)

   Amuse : Platitudes
   Compel : Threats
   Deter : Tidings
   Batter : Insults
   Exercise : Antics

10. Quiet : Sound :: Darkness : (.95; -.11)

    Cellar
    Sun
    Noise
    Stillness
    Light

11. Tall : Enormous :: Short : (.95; .24)

    Tiny
    Medium
    Gigantic

Ravenous
Hungry

12. Intransigent : Flexibility :: (.08; .37)

Transient : Mobility
Disinterested : Partisanship
Dissimilar : Variation
Progressive : Transition
Ineluctable : Modality

13. Deference : Respect :: (.22; .15)

Admiration : Jealousy
Condescension : Hatred
Affection : Love
Pretence : Truth
Gratitude : Charity

14. Tragedy : Drama :: (.37; .06)

Farce : Actor
Cartoon : Film
Prosody : Poem
Accident : Ambulance
Epigram : Anecdote

15. Butterfly : Caterpillar :: Frog : (.91;.41)

Fish
Amphibian
Frog
Toad
Tadpole

16. Century : 100 :: Decade : (.97; .28)

1000
10
Millennium
Score
1/10

17. Square : Cube :: Circle : (.92; .08)

Rhombus
Corner
Sphere
Cylinder
Ellipsoid

18. Prison : Criminals :: Orchard : (.58; .20)

Fruit
Trees
Pickers
Grass
Oranges

19. Success : Elation :: Failure : (.85; .25)

Fear
Euphoria
Contagion
Hospitality
Depression

20. Lion : Mammal :: Bumblebee : (.95;.22)

Amphibian
Hive
Aphid
Insect
Winged

22. Tenet : Theologian :: (.43; .34)

Predecessor : Heir
Hypothesis : Biologist
Recluse : Rivalry
Arrogance : Persecution
Guitarist : Rock Band

23. Walk : Legs :: (.66; .07)

Blink : Eyes
Chew : Mouth
Dress : Hem
Cover : Book
Grind : Nose

24. Whisper : Shout :: Walk : (.91; -.07)

Command
Fly
Gallop
Tiptoe
Run

25. Good : Better :: Bad : (.88; .09)

Worst
Worse
Better
Badder
Best

26. Circle : Sphere :: Square : (.95; .20)

Globe
Oblong
Cube
Diamond
Box

27. Scintillating : Dullness :: (.68; .42)

Erudite : Wisdom
Desultory : Error
Boisterous : Calm
Cautious : Restraint
Exalted : Elevation

28. Much : More :: Little : (.88; .10)

All
Less
A Lot
Big
Small

29. Week : Fortnight :: Solos : (.38; .45)

Choir
Singers
Trio
Twins

Duet

30. Morbid : Unfavorable :: (.69; .35)

Reputable : Favorable
Maternal : Unfavorable
Disputations : Favorable
Vigilant : Unfavorable
Lax : Favorable

31. Flower : Bouquet :: Link : (.89; .31)

Connect
Event
Chain
Joining
Braid

32. Rite : Ceremony :: (.37; .23)

Magnitude : Size
Affliction : Blessing
Clamor : Silence
Pall : Clarity
Agitation : Calm

33. Hackneyed : Freshness :: (.26; .16)

Stale : Porosity
Facile : Delicacy
Ponderous : Lightness
Central : Vitality
Relevant : Pertinence

34. Inflate : Bigger :: (.92; .24)

Revere : Lower
Elongate : Shorter
Fluctuate : Longer
Mediate : Higher
Diminish : Smaller

35. Author : Literate :: (.57; .49)

Cynic : Gullible

Hothead : Prudent
Saint : Notorious
Judge : Impartial
Doctor : Fallible

36. Attenuate : Signal :: (.11; .19)

Exacerbate : Problem
Modify : Accent
Dampen : Enthusiam
Elongate : Line
Dramatize : Play

37. Humdrum : Bore :: (.23; .49)

Grim : Amuse
Nutritious : Sicken
Stodgy : Excite
Heartrending : Move
Pending : Worry

38. Reinforce : Stronger :: (.94; .25)

Abound : Lesser
Dismantle : Longer
Wilt : Higher
Shirk : Greater
Erode : Weaker

39. Braggart : Modesty :: (.37; .47)

Fledgling : Experience
Embezzler : Greed
Wallflower : Timidity
Invalid : Malady
Candidate : Ambition

40. Ski : Snow :: (.91; .18)

Drive : Car
Gold : Putt
Dance : Step
Skate : Ice
Ride : Horse

41. Verify : True :: (.92; .29)

Signify : Cheap
Purify : Clean
Terrify : Confident
Ratify : Angry
Mortify : Relaxed

42. Tarantula : Spider :: (.35; .39)

Mare : Stallion
Milk : Cow
Fly : Parasite
Sheep : Grass
Drone : Bee

43. Prosaic : Mundane :: (.25; -.21)

Obdurate : Foolish
Ascetic : Austere
Clamorous : Captive
Loquacious : Taciturn
Peremptory : Spontaneous

44. Doctor : Hospital :: (.91; .20)

Sports Fan : Stadium
Cow : Farm
Professor : College
Criminal : Jail
Food : Grocery Store

45. Cub : Bear :: (.97; .34)

Piano : Orchestra
Puppy : Dog
Cat : Kitten
Eagle : Predator
Fork : Utensil

46. Salacious : Wholesome :: (.34; .48)

Religious : Private
Expensive : Profligate
Conservative : Stoic
Mendacious : Truthful
Fulsome : Generous

47. Elected : Inauguration :: (.52; .57)

Enrolled : Graduation
Condemned : Execution
Chosen : Selection
Gathered : Exhibition
Appointed : Interview

48. Dividend : Stockholder :: (.14; .32)

Patent : Inventor
Royalty : Author
Wage : Employer
Interest : Banker
Investment : Investor

49. Mince : Walk :: (.45; .17)

Bang : Sound
Wave : Gesture
Waltz : Dance
Simpler : Smile
Hike : Run

50. Disinterested : Unbiased :: (.18; .32)

Indulgent : Intolerant
Exhausted : Energetic
Languid : Lethargic
Unconcerned : Involved
Profilgate : Flippant

51. Temper : Hard :: (.08; .17)

Mitigate : Severe
Provoke : Angry
Endorse : Tough
Infer : Certain
Scrutinize : Clear

52. Stanza : Poem :: (.83; .32)

Chaper : Novel
Prose : Verse
Stave : Music

Song : Chorus
Overture : Opera

53. Stoic : Fortitude :: (.32; .12)

Benefactor : Generosity
Heretic : Faith
Eccentric : Ineptitude
Miser : Charity
Soldier : Bravery

54. Ambivalent : Certain :: (.11; -.18)

Indifferent : Biased
Furtive : Open
Impecunious : Voracious
Discreet : Careful
Munificent : Generous

55. Allay : Suspicion :: (.29; .21)

Tend : Plant
Impede : Anger
Calm : Fear
Fell : Tree
Exacerbate : Worry

56. Primitive : Sophisticate :: (.14; .09)

Employer : Superior
Socialite : Recluse
Tyro : Expert
Native : Inhabitant
Applicant : Member

57. Authoritarian : Lenient :: (.38; .47)

Philanthropist : Generous
Virtuoso : Glamorous
Hedonist : Indulgent
Servant : Servile
Miser : Charitable

58. Perennial : Ephemeral :: (.45; .23)

Volatile : Evanescent
Mature : Ripe
Diurnal : Annual
Permanent : Temporary
Majestic : Mean

59. Needle : Pine :: (.60; .40)

Pin : Cloth
Trunk : Tree
Flower : Leaf
Spine : Cactus
Stalk : Root

60. Anecdote : Story :: (.38; .40)

Ballad : Song
Novel : Chapter
Limerick : Poem
Prose : Poetry
Overture : Opera

Appendix B – Manipulation Check

1) Answer this question as if you couldn't decide between answer choice 'A' and answer choice 'B', but felt slightly more certain of answer choice 'B'.
    a.
    b.
    c.
    d.
    e.
2) Answer this question as if you were fairly certain that answer choice 'C' was correct.
    a.
    b.
    c.
    d.
    e.
3) Answer this question as if you had no idea as to which answer choice is correct.
    a.
    b.
    c.
    d.
    e.
4) Answer this question as if you were certain that answer choice 'D' and answer choice 'E' were incorrect, but had no strong idea of which of the remaining was correct.
    a.
    b.
    c.
    d.
    e.
5) Answer this question as if you were fairly certain answer choice 'A' was correct, but had some thoughts that there was a chance that answer choice 'B' was correct.
    a.
    b.
    c.
    d.
    e.

Appendix C – Generalized Anxiety Items

From "The Items in Each of the Preliminary IPIP Scales
Measuring Constructs Similar to Those in the NEO-PI-R"
http://ipip.ori.org/newNEOKey.htm#Anxiety

In general, I…

N1: ANXIETY (Alpha = .83)

+ keyed Worry about things.
Fear for the worst.
Am afraid of many things.
Get stressed out easily.
Get caught up in my problems.

– keyed Am not easily bothered by things.
Am relaxed most of the time.
Am not easily disturbed by events.
Don't worry about things that have already happened.
Adapt easily to new situations.

1- Strongly Disagree
2- Disagree
3- Neither Agree or Disagree
4- Agree
5- Strongly Agree

Appendix D – Test Anxiety Items from Taylor and Deane (2002)

1) During tests I feel very tense.
2) I wish examinations did not bother me so much.
3) I seem to defeat myself while working on important tests.
4) I feel very panicky when I take an important test.
5) During examinations I get so nervous I forget facts I really know.

1- Strongly Disagree
2- Disagree
3- Neither Agree or Disagree
4- Agree
5- Strongly Agree

Appendix E – Risk Taking Items

From "The Items in the 15 Preliminary IPIP Scales Measuring Constructs Similar to Those in the Jackson Personality Inventory (JPI-R)" http://ipip.ori.org/newJPI-RKey.htm#Risk-Taking

Generally, I…

RISK-TAKING (JPI: Risk Taking [Rkt]) [.78]

+ keyed	Enjoy being reckless.
	Take risks.
	Seek danger.
	Know how to get around the rules.
	Am willing to try anything once.
	Seek adventure.
– keyed	Would never go hang-gliding or bungee-jumping.
	Would never make a high risk investment.
	Stick to the rules.
	Avoid dangerous situations.

1- Strongly Disagree
2- Disagree
3- Neither Agree or Disagree
4- Agree
5- Strongly Agree

Appendix F – Cautiousness Items

From "The Items in Each of the Preliminary IPIP Scales Measuring Constructs Similar to Those in the 30 NEO-PI-R Facet Scales" http://ipip.ori.org/newNEOFacetsKey.htm

Generally, I…

C6: CAUTIOUSNESS (.76)

+ keyed Avoid mistakes.
Choose my words with care.
Stick to my chosen path.

– keyed Jump into things without thinking.
Make rash decisions.
Like to act on a whim.
Rush into things.
Do crazy things.
Act without thinking.
Often make last-minute plans.

1- Strongly Disagree
2- Disagree
3- Neither Agree or Disagree
4- Agree
5- Strongly Agree

Appendix G – Self-Efficacy Items

From "The Items in Each of the Preliminary IPIP Scales Measuring Constructs Similar to Those in the 30 NEO-PI-R Facet Scales" http://ipip.ori.org/newNEOFacetsKey.htm

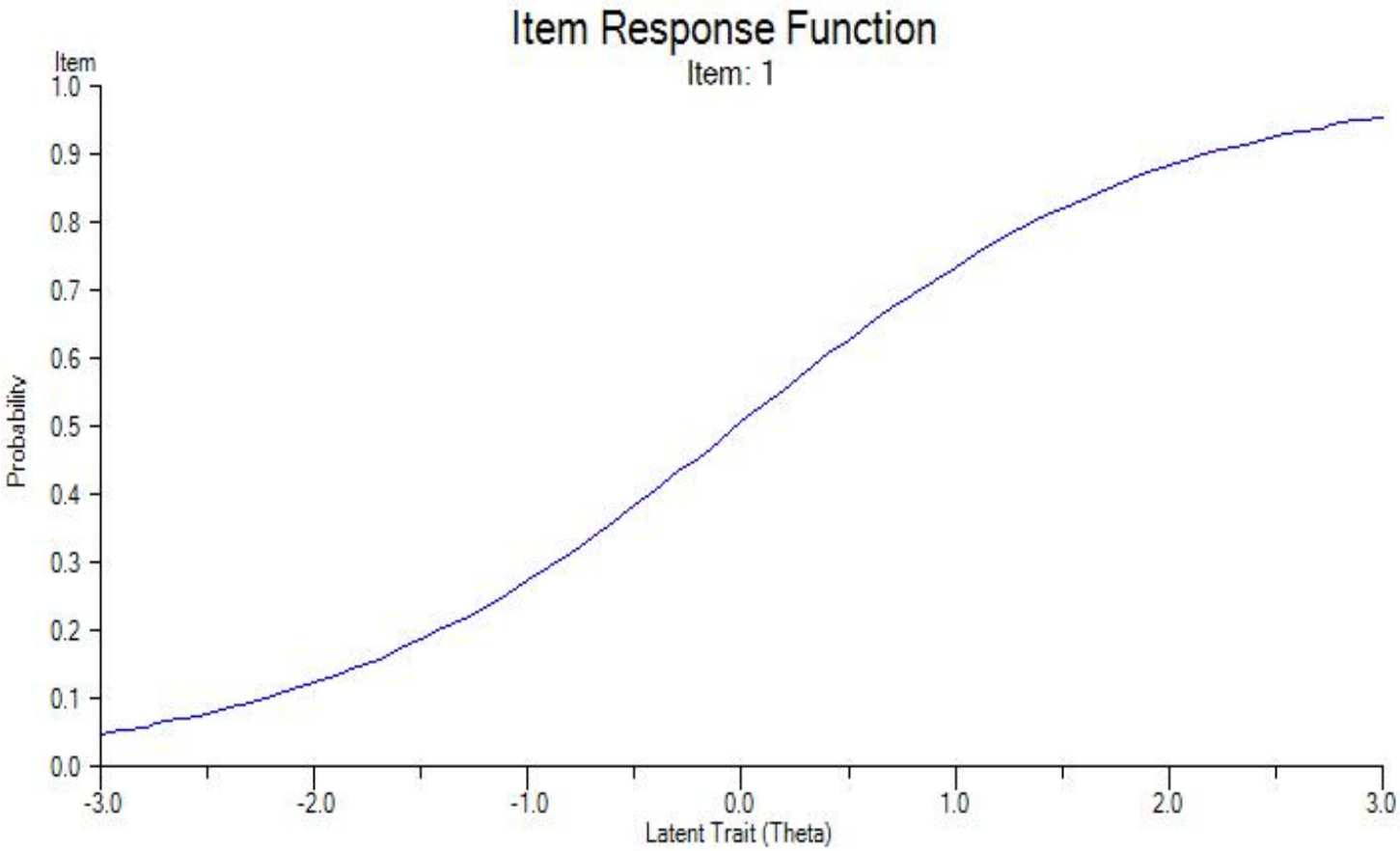
When taking standardized tests, I…

C1: SELF-EFFICACY (.78)

+ keyed	Complete tasks successfully.
	Excel in what I do.
	Handle tasks smoothly.
	Am sure of my ground.
	Come up with good solutions.
	Know how to get things done.
– keyed	Misjudge situations.
	Don't understand things.
	Have little to contribute.
	Don't see the consequences of things.

1- Strongly Disagree
2- Disagree
3- Neither Agree or Disagree
4- Agree
5- Strongly Agree

Figure 1 – The Item Characteristic Curve



NOTE: For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Figure 2 – Item Characteristic Curve for the Nominal Response Model

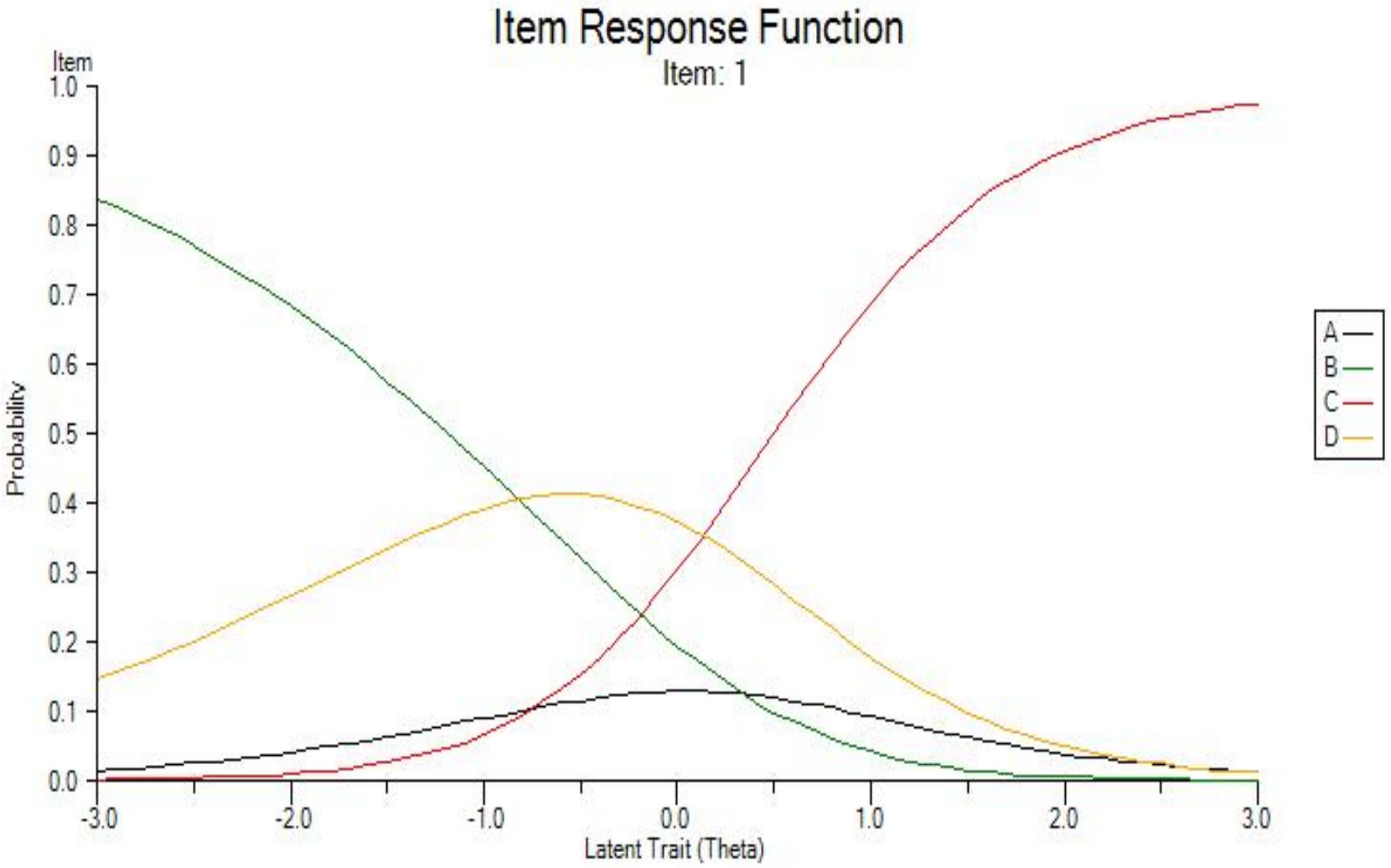


Figure 3 – Hierarchical Framework for Modeling Response Time and Response
(van der Linden, 2007, p295)

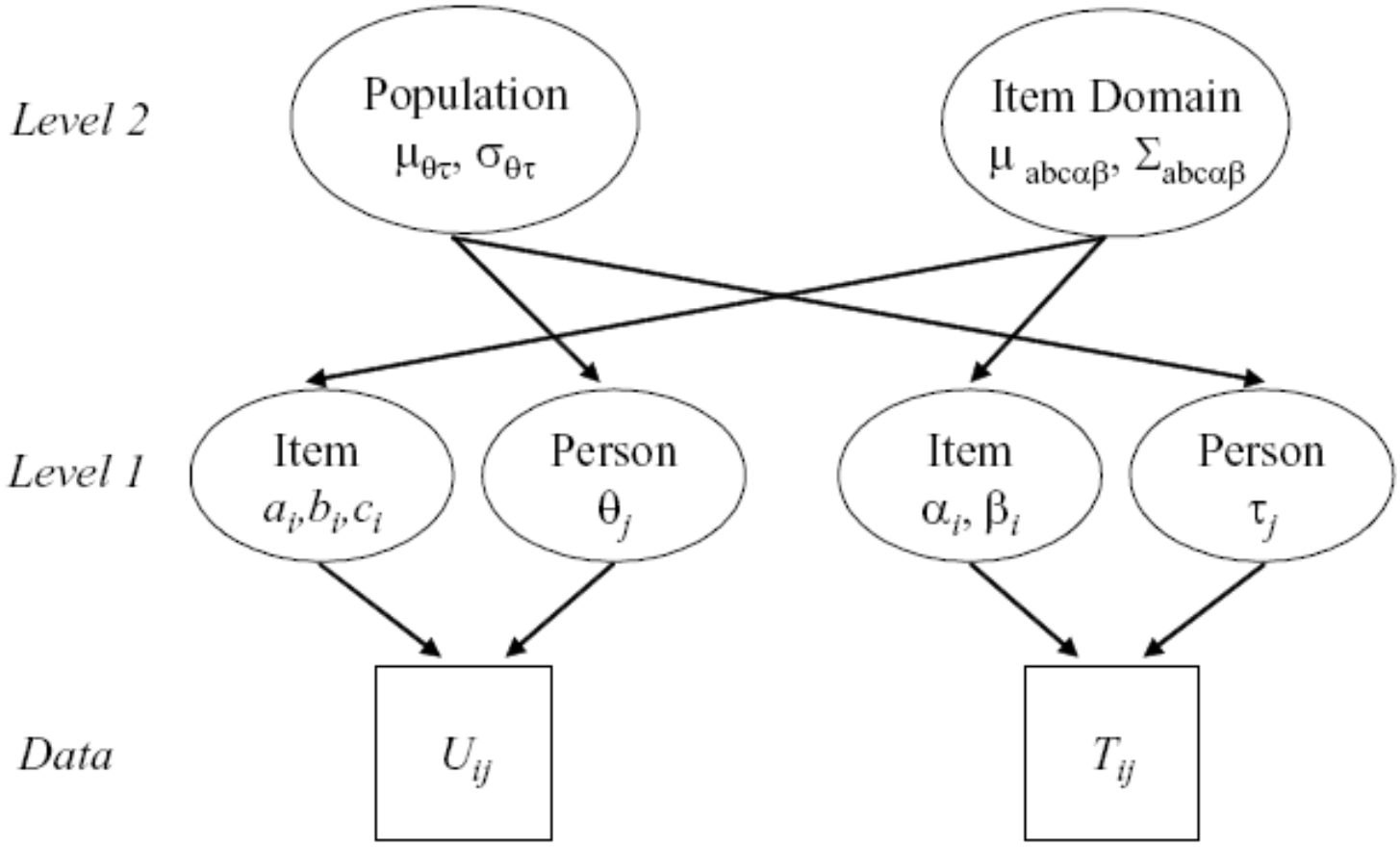


Figure 4: Test reliability as a function of test length.

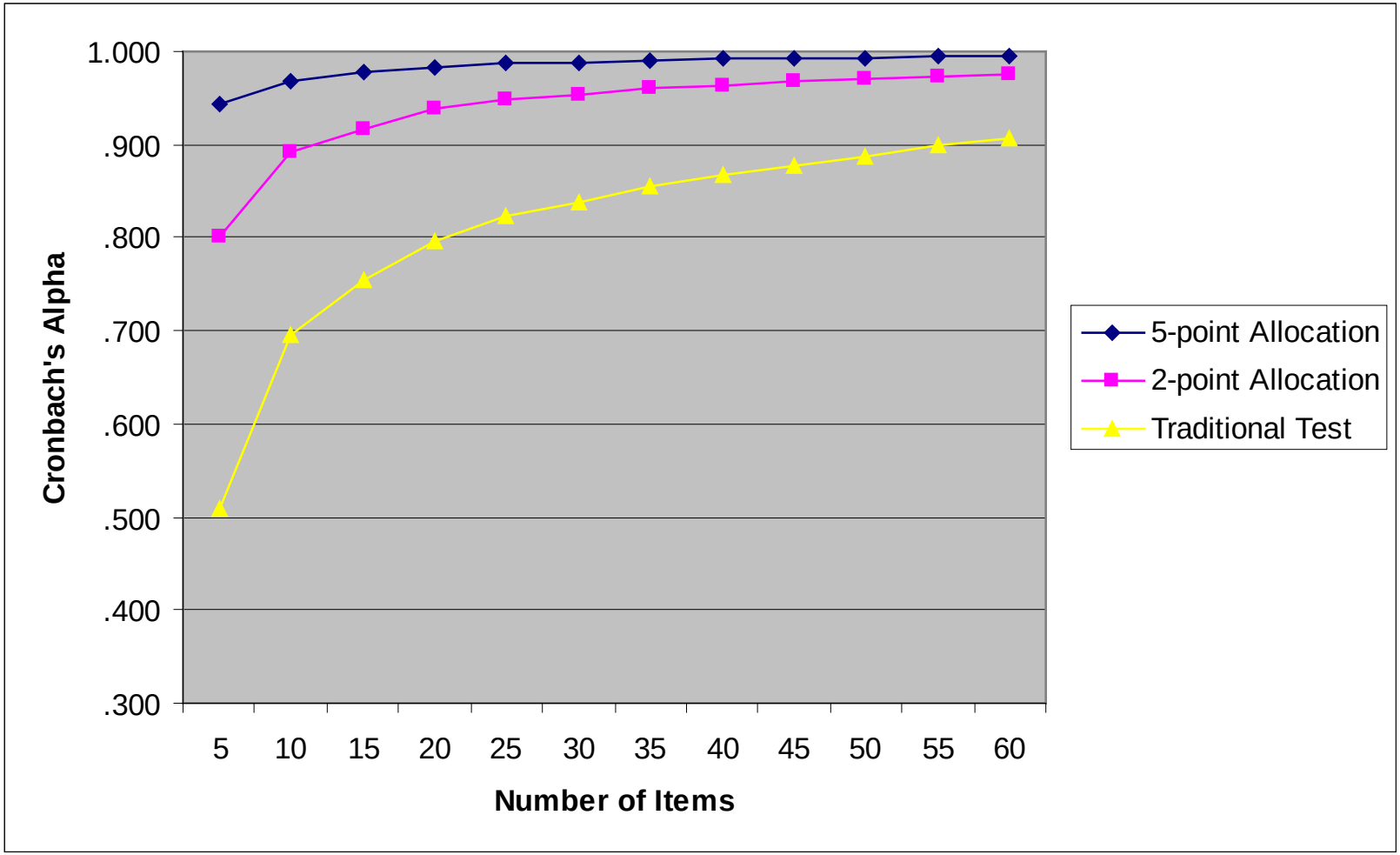


Figure 5: Benefit of confidence testing as a function of test length.

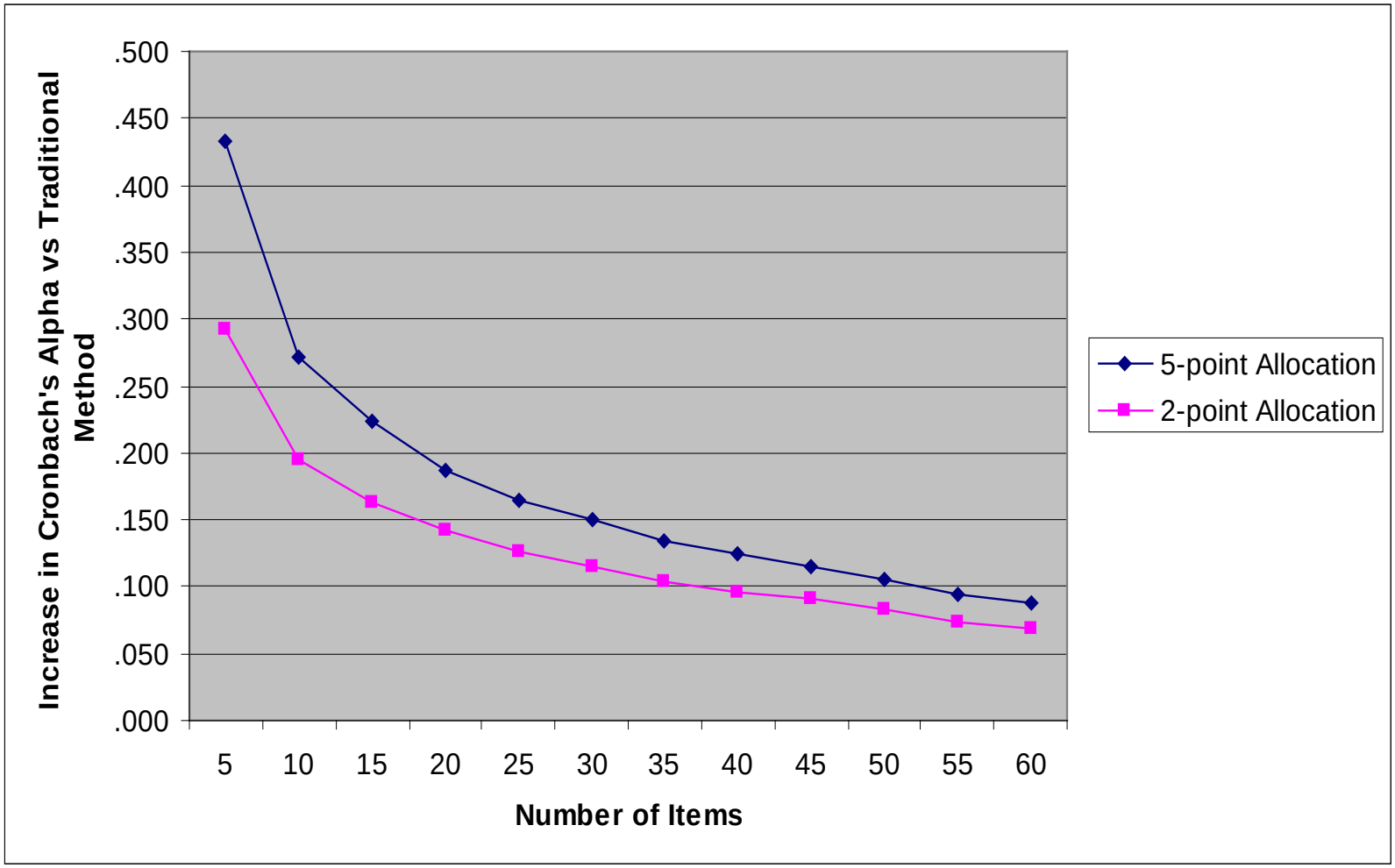


Figure 6: Test reliability as a function of test length and difficulty.

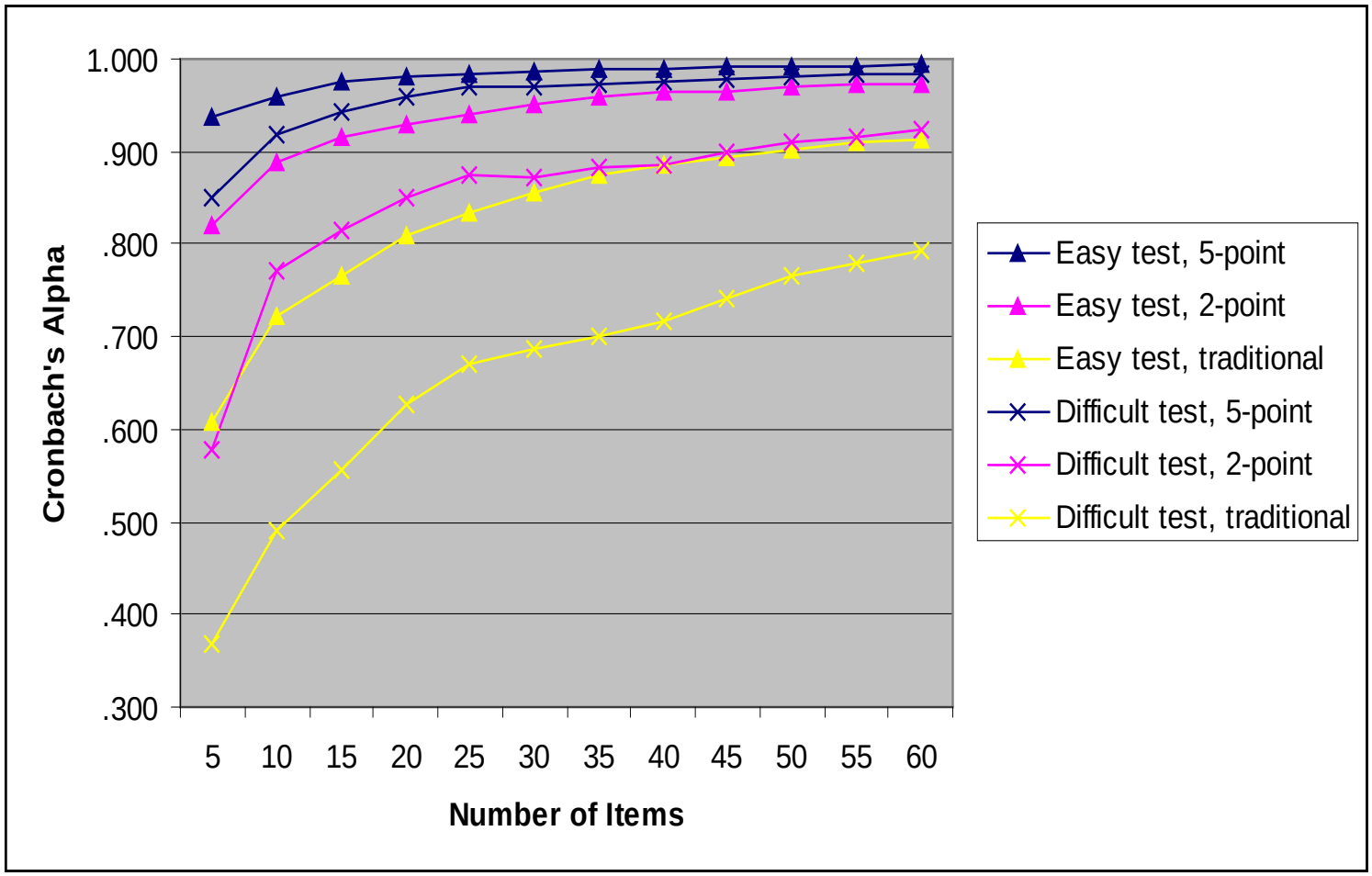


Figure 7: Benefits of confidence testing as a product of test length and difficulty.

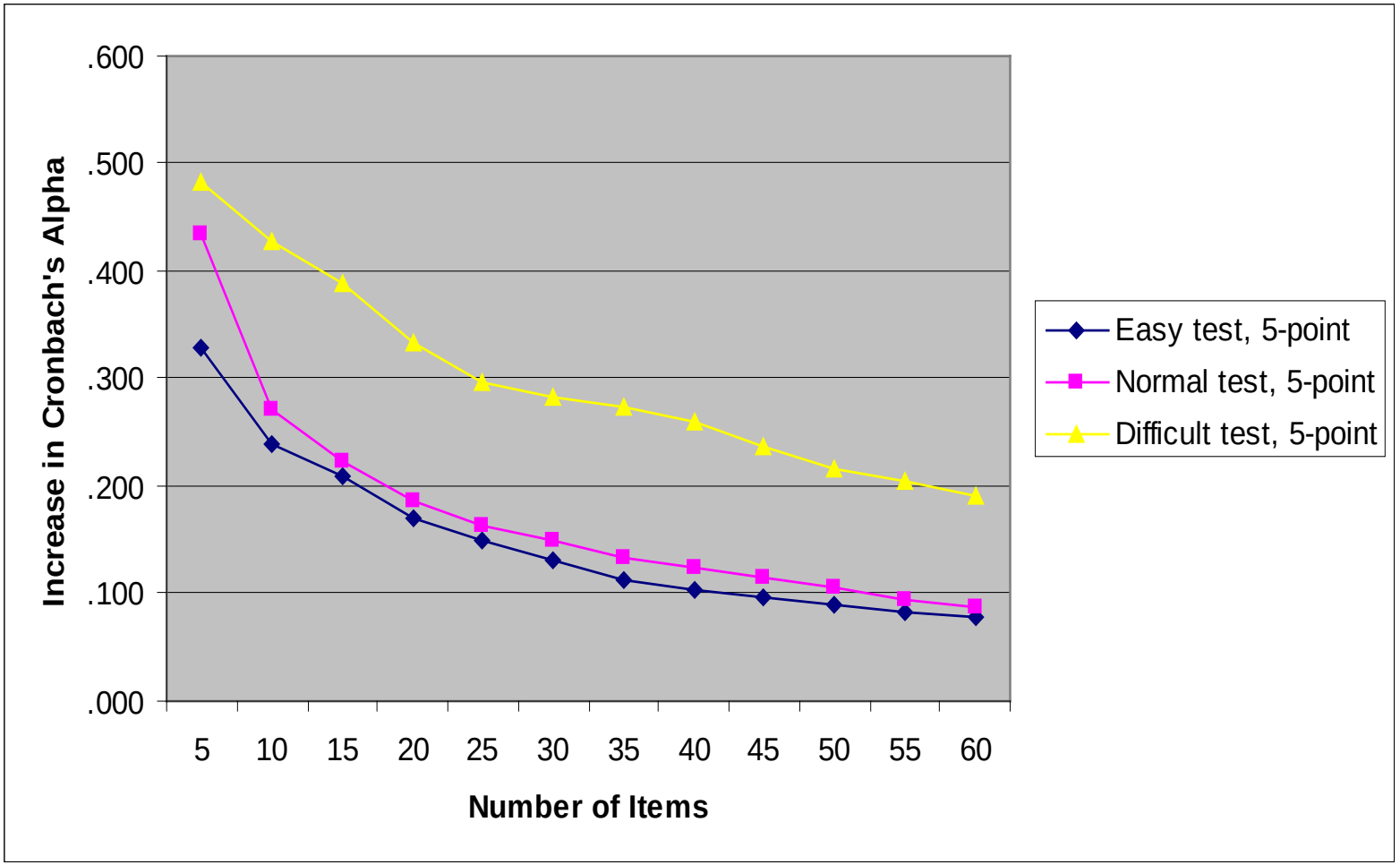


Figure 8: Test reliability as a function of test length and discrimination.

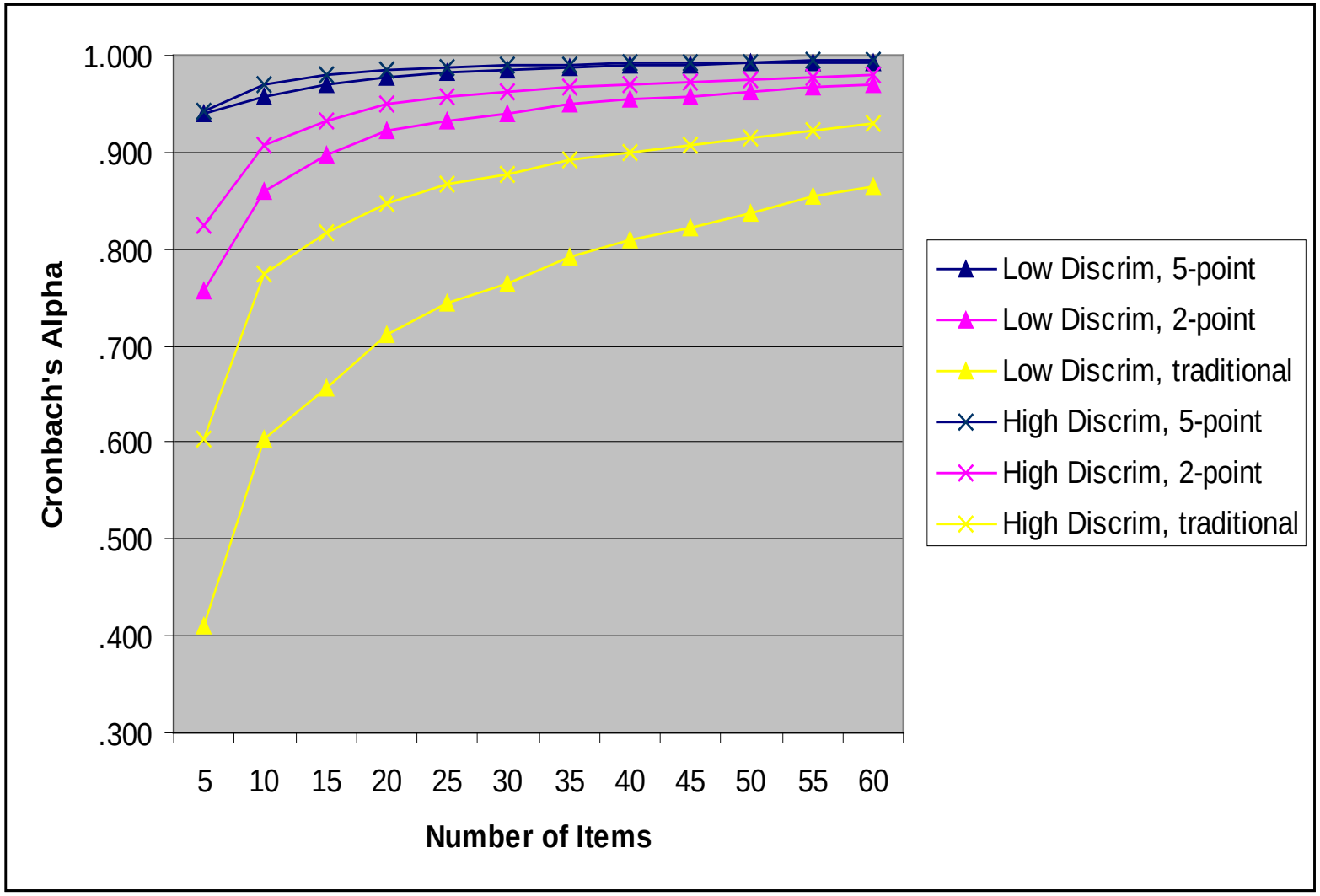


Figure 9: Benefits of confidence testing as a product of test length and discrimination.

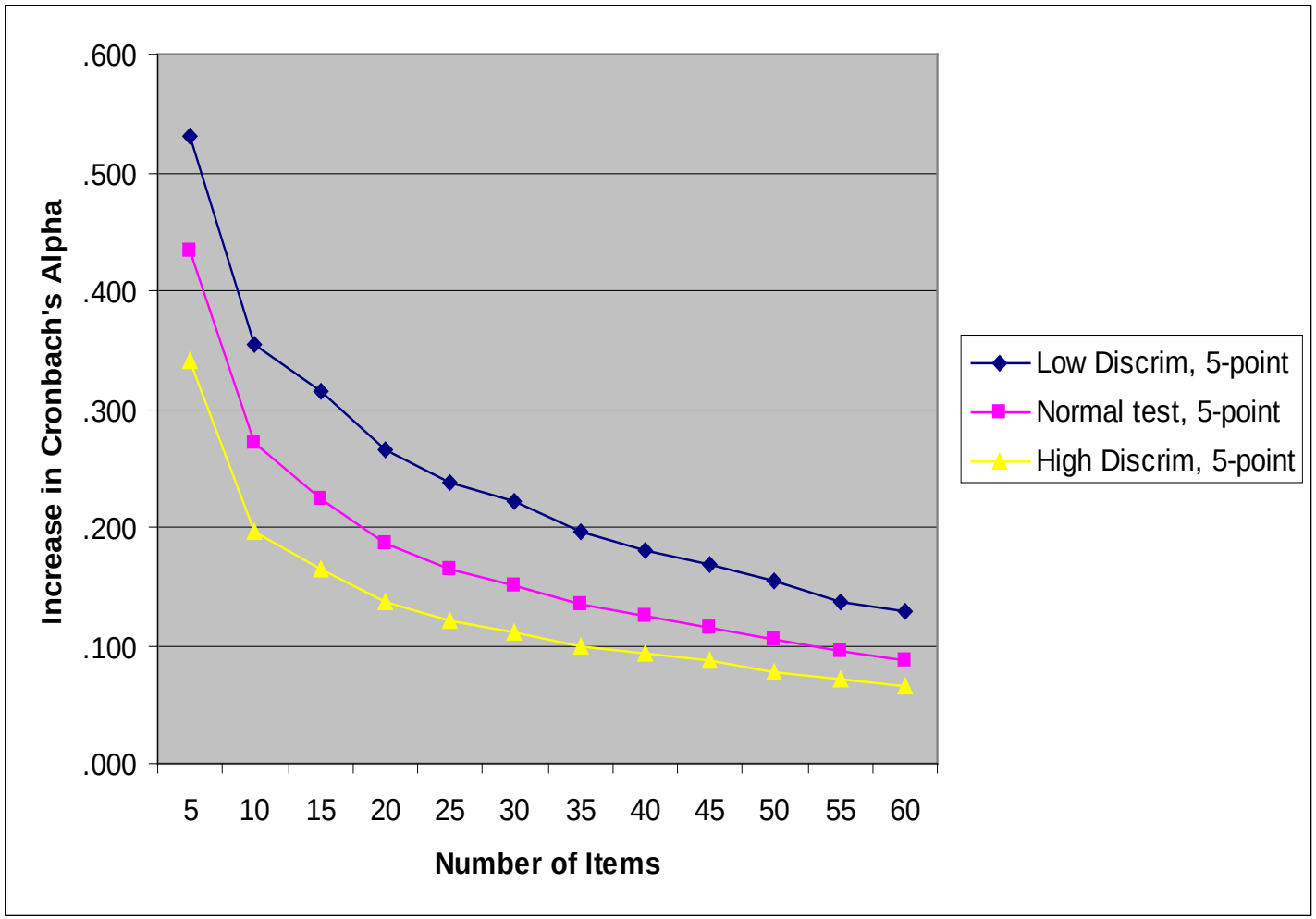


Figure 10 – Distribution of confidence scores in 5-point confidence condition.

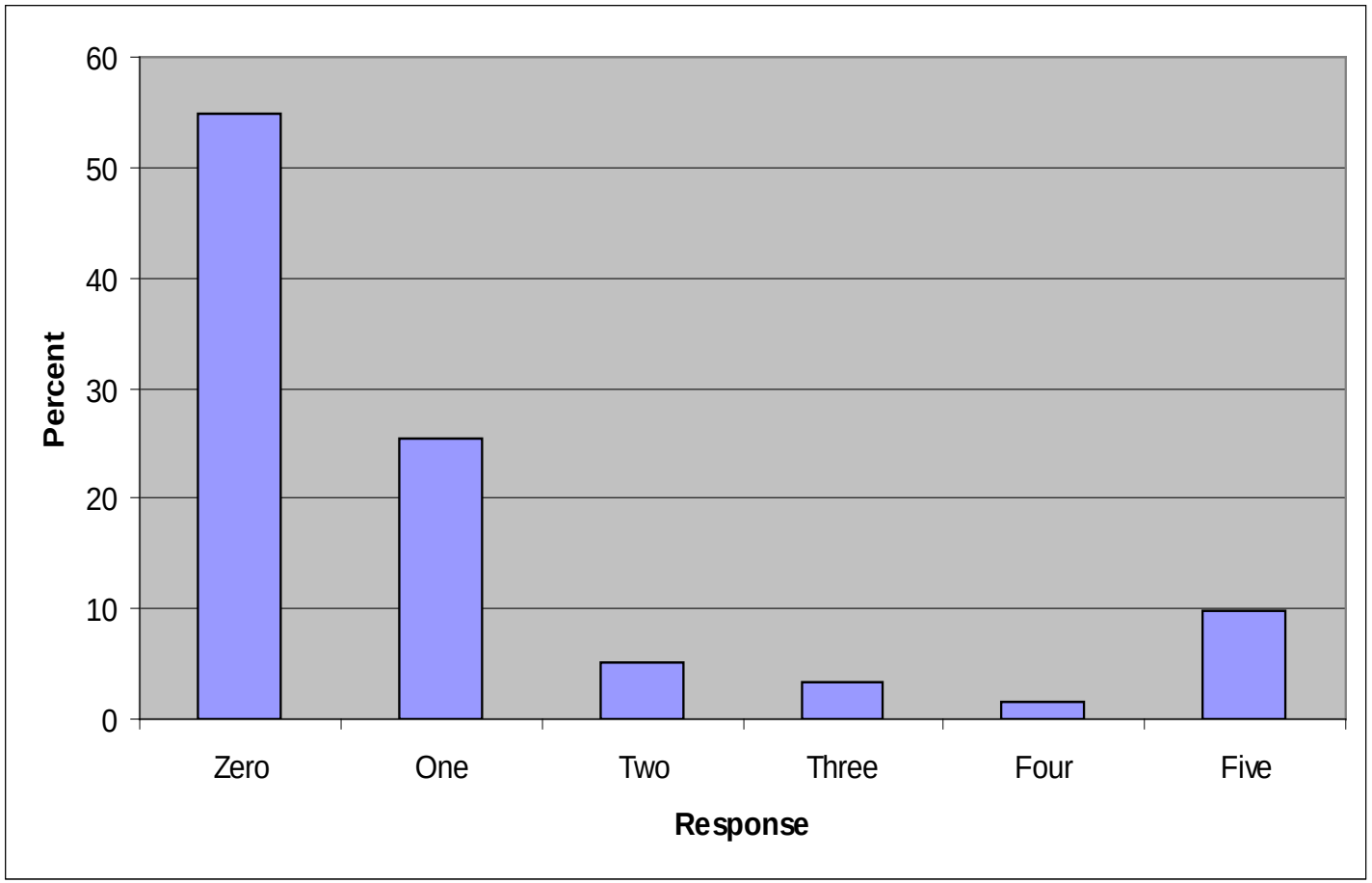


Figure 11 – Distribution of confidence scores in generalized confidence condition.

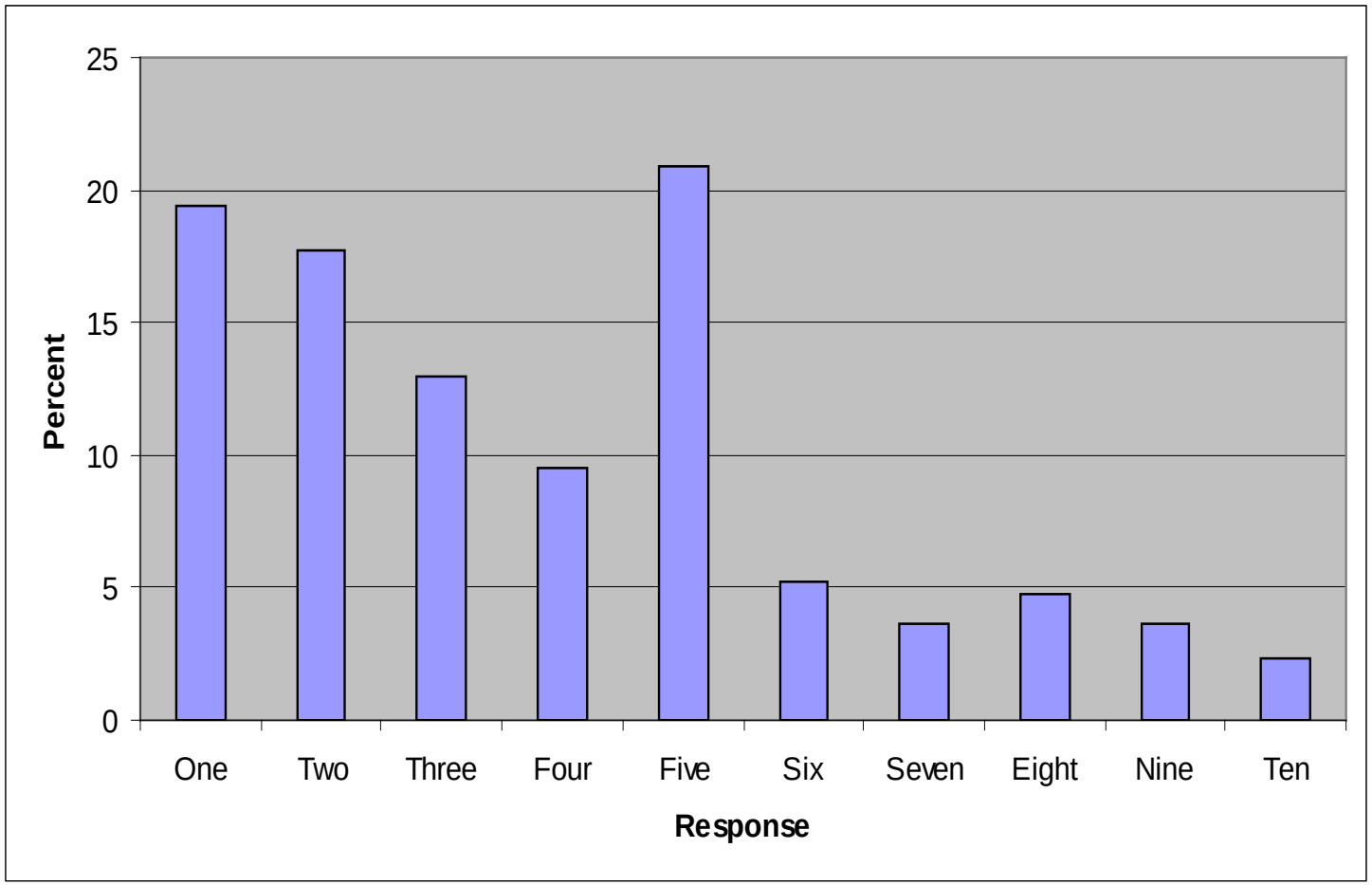


Figure 12 – Reliability by Test Length and Condition.

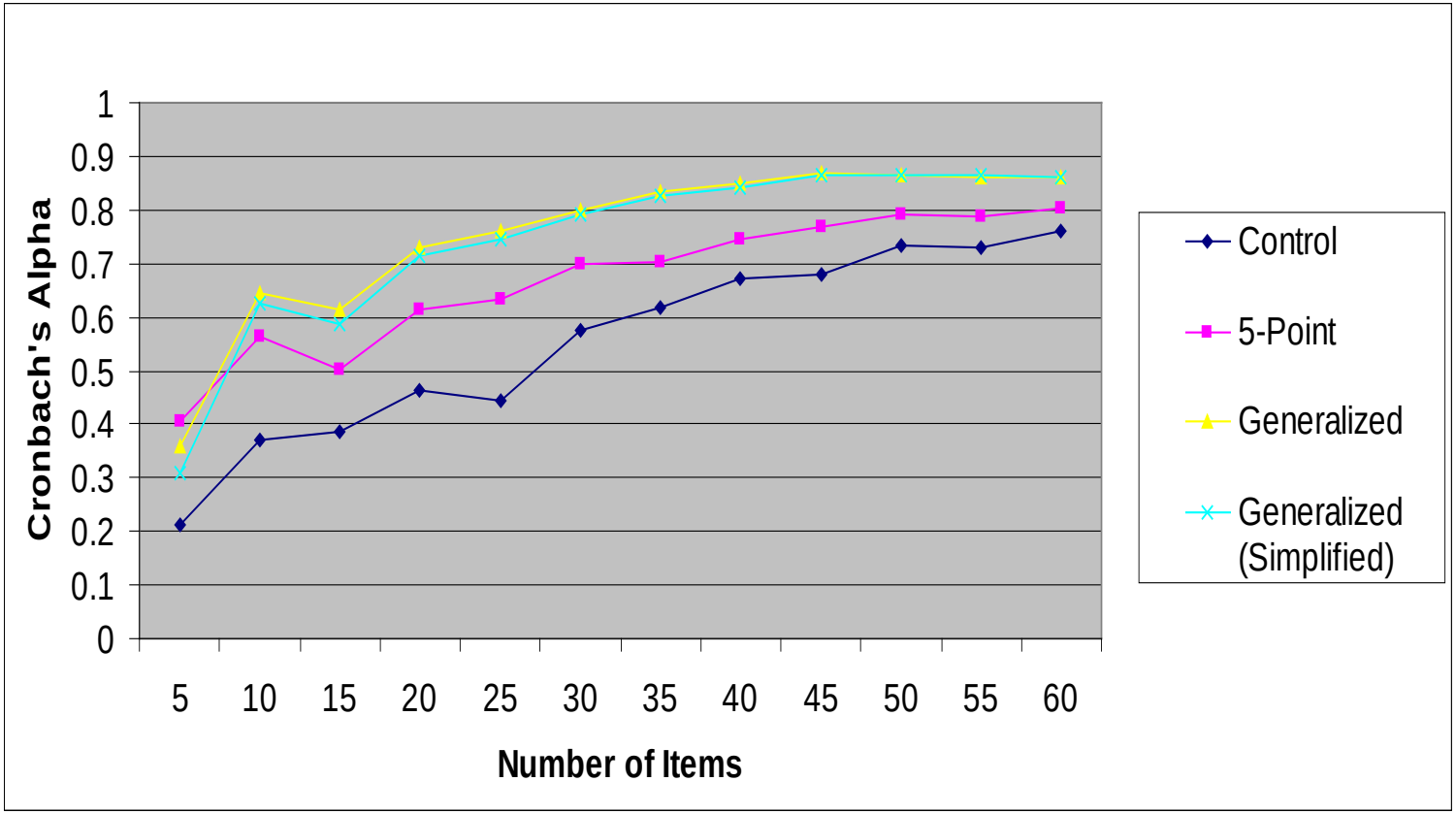


Figure 13 – Reliability Benefits of Confidence Testing Relative to Control.

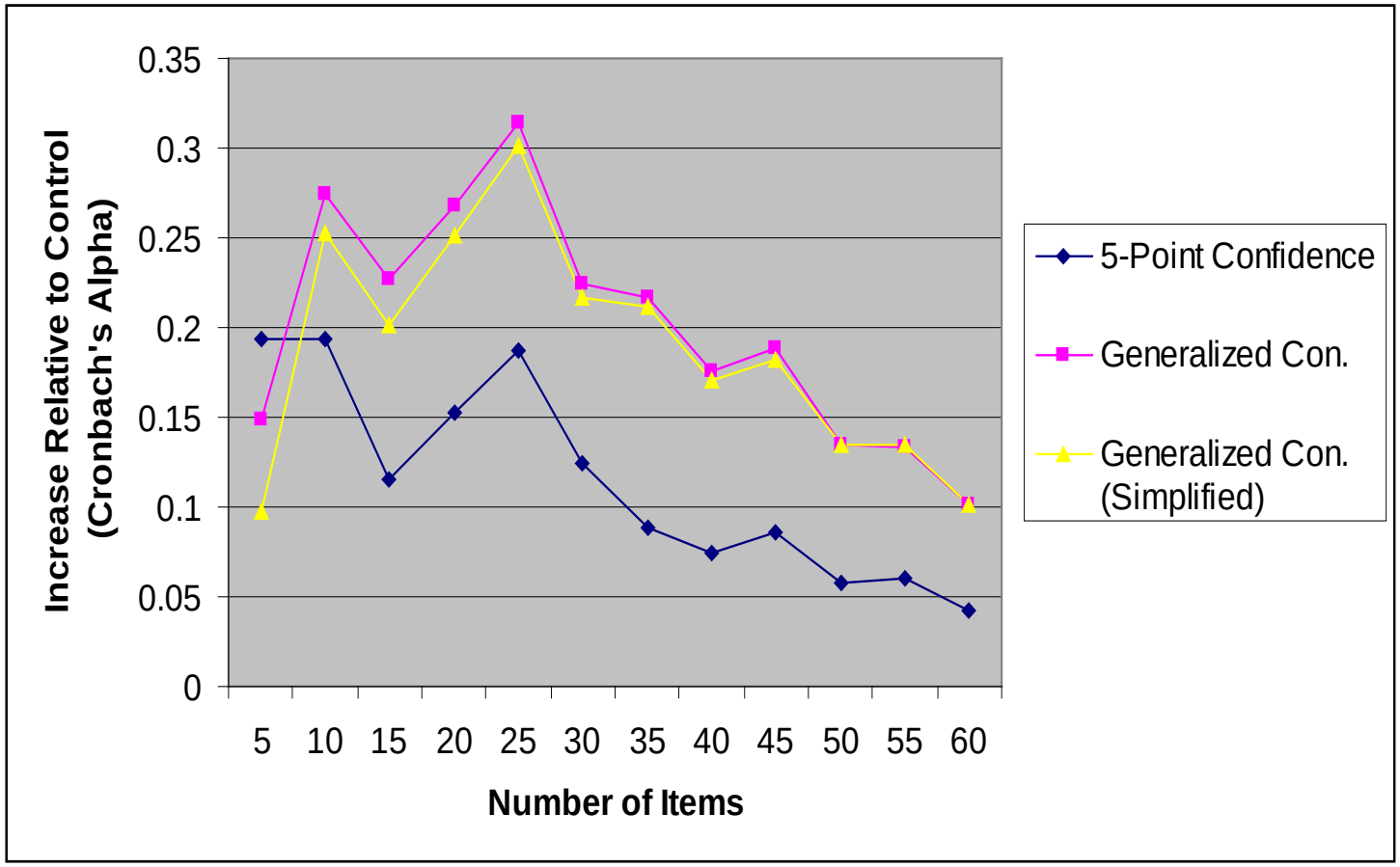


Figure 14 – Validity (College GPA) by Length.

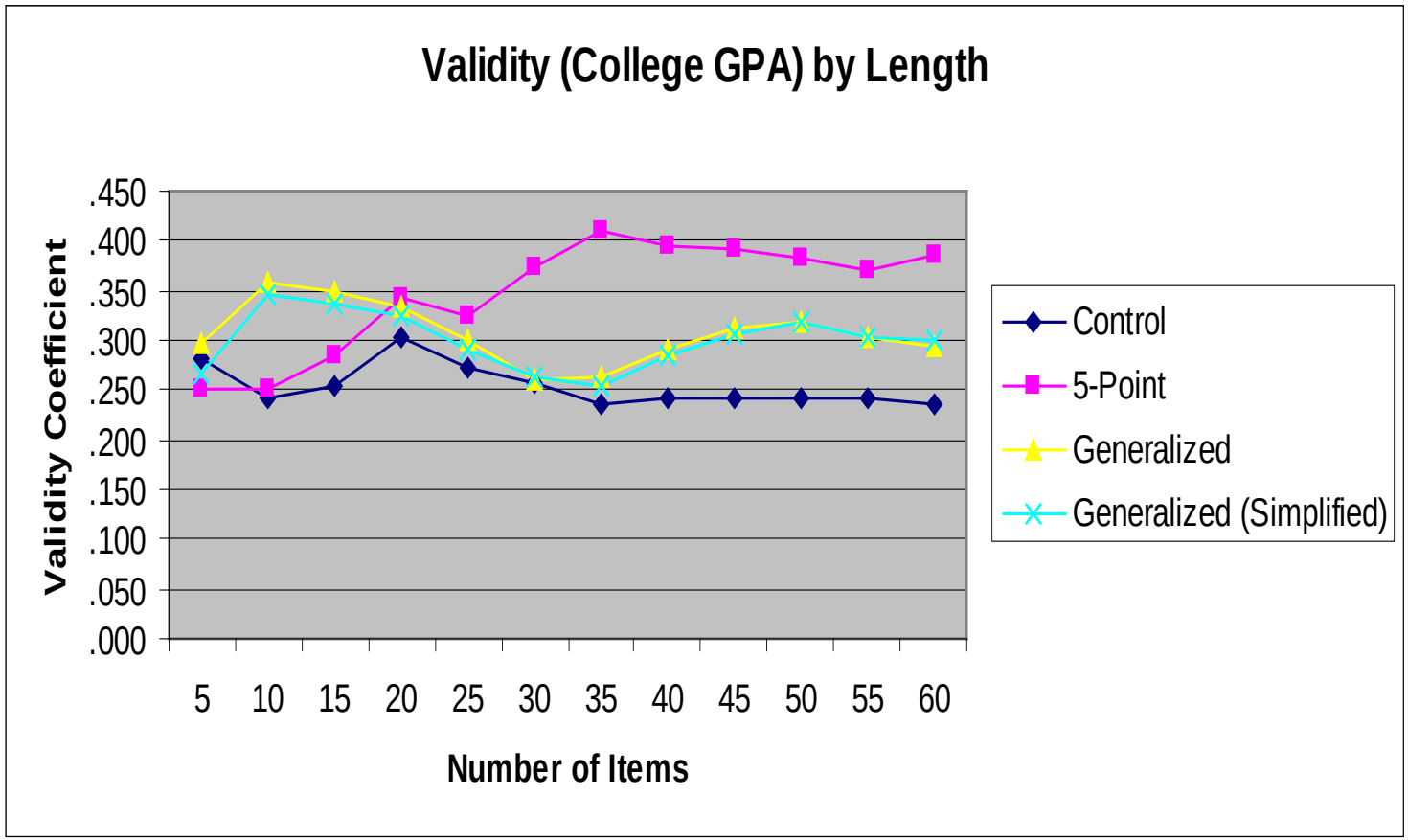


Figure 15 – Validity (ACT score) by Length.

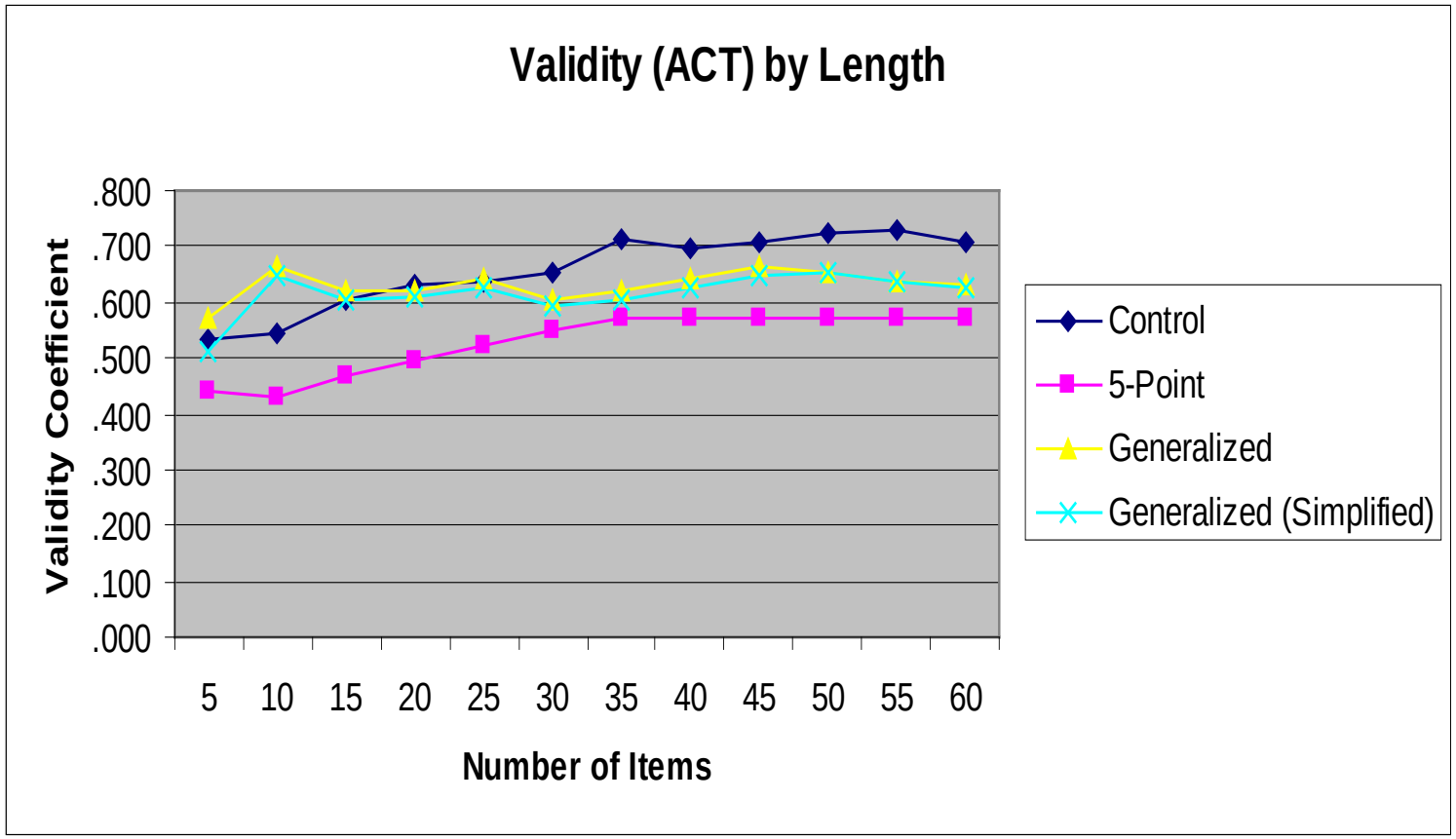


Figure 16 – Average Time of Test by Length of Test, by Condition.

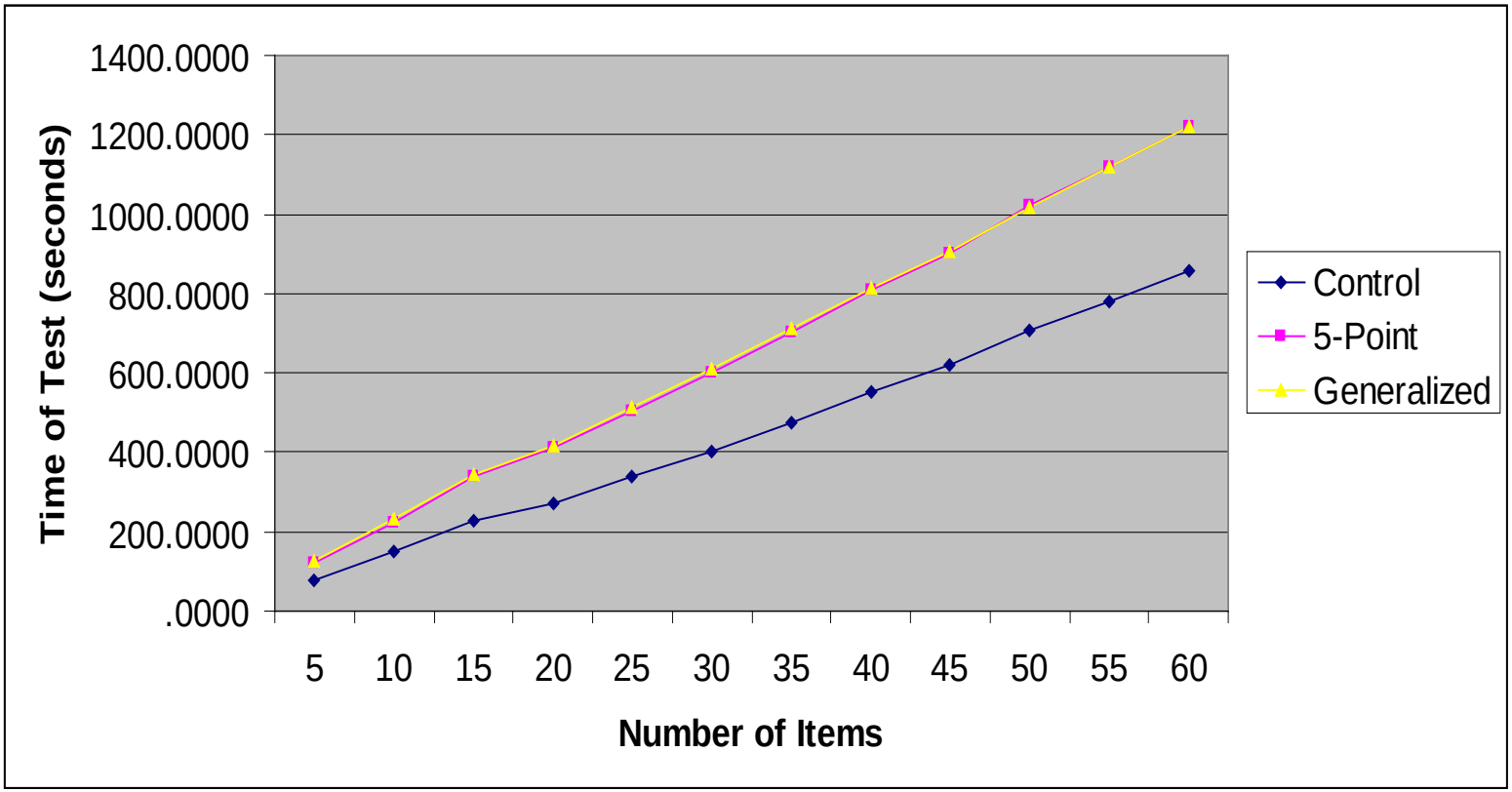


Figure 17 – Comparison of Item Discrimination by Test and Item Difficulty

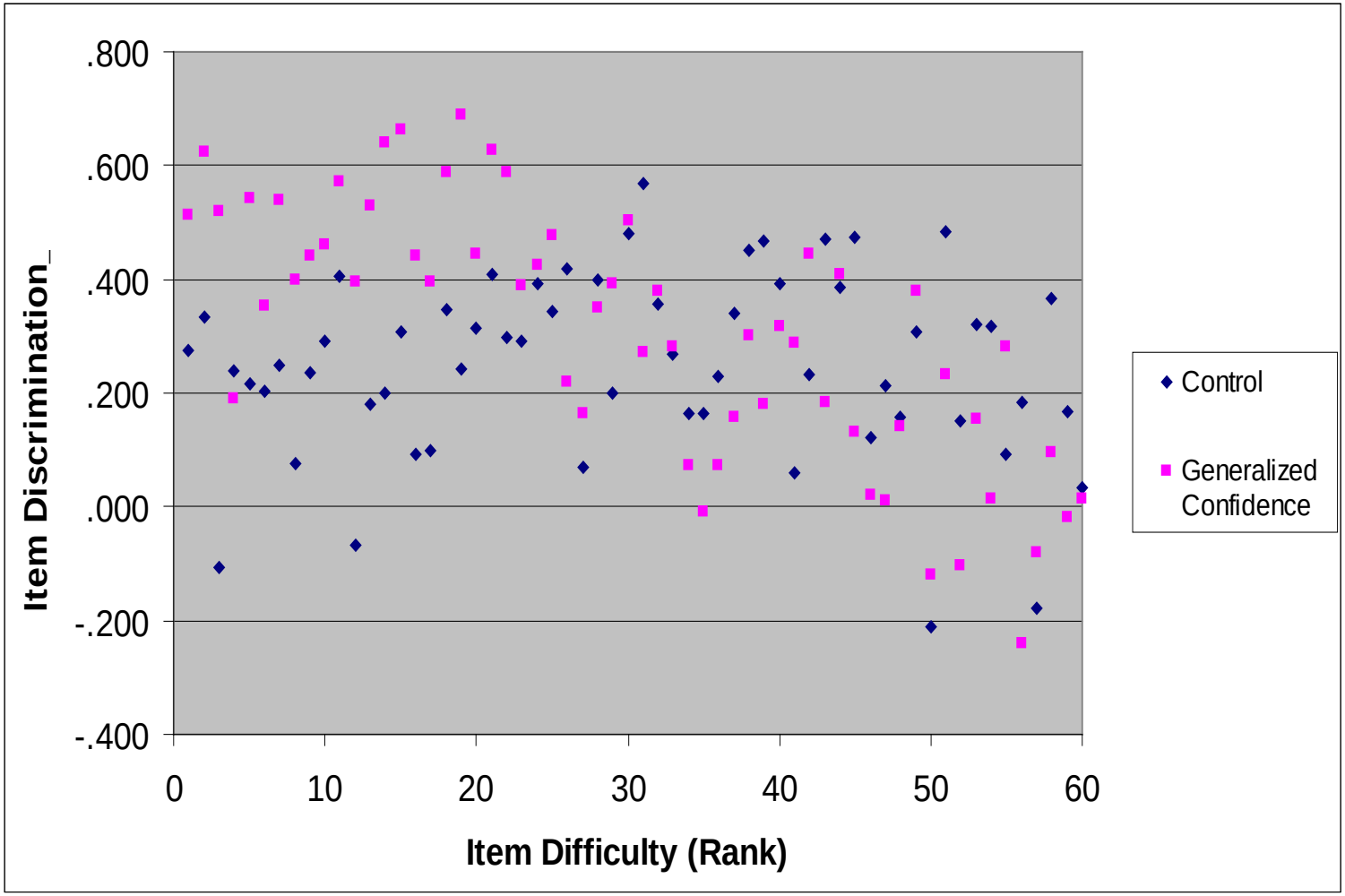


Figure 18 – Reliability by Test Difficulty.

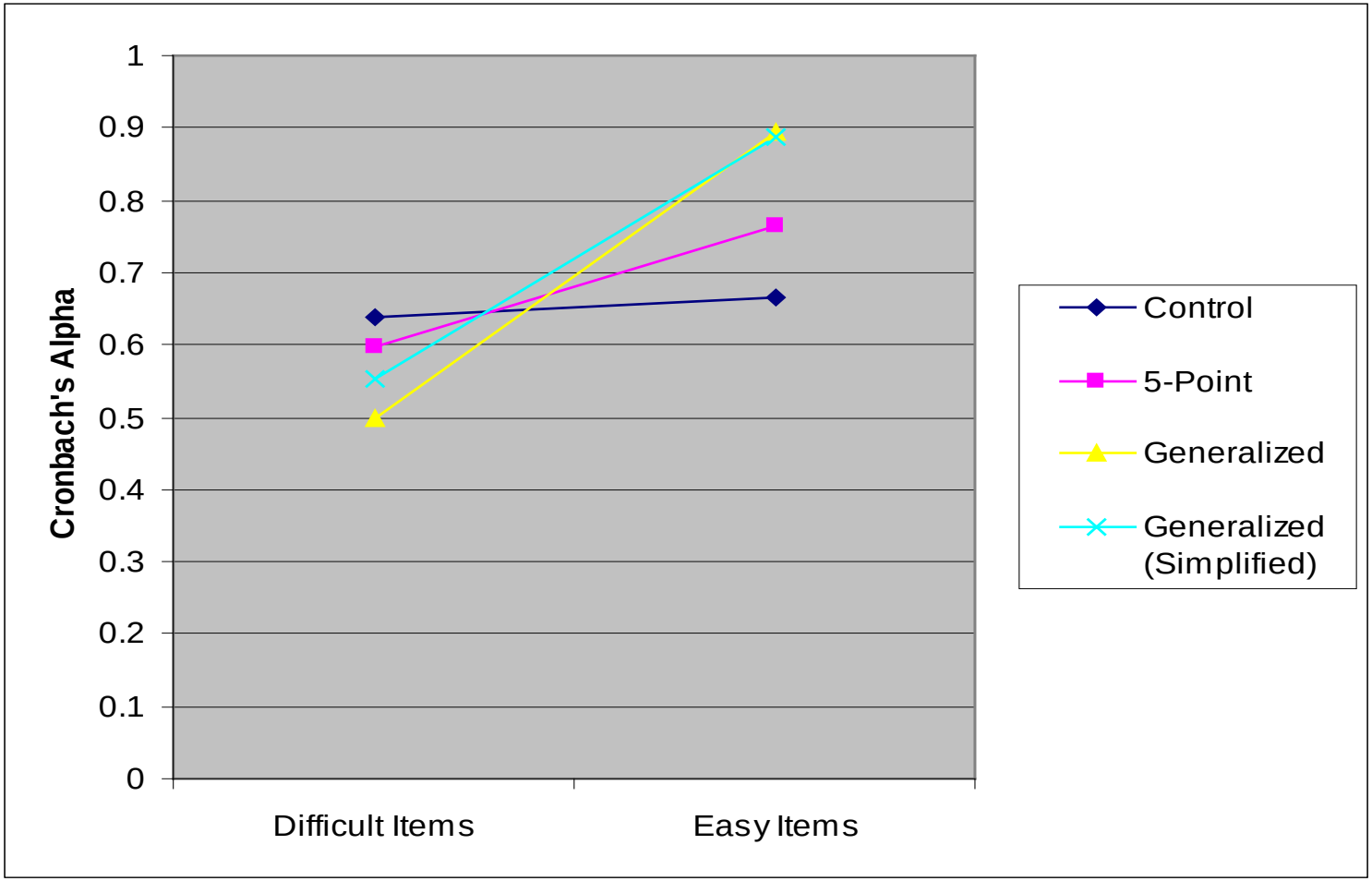


Figure 19 – Reliability by Test Discrimination.

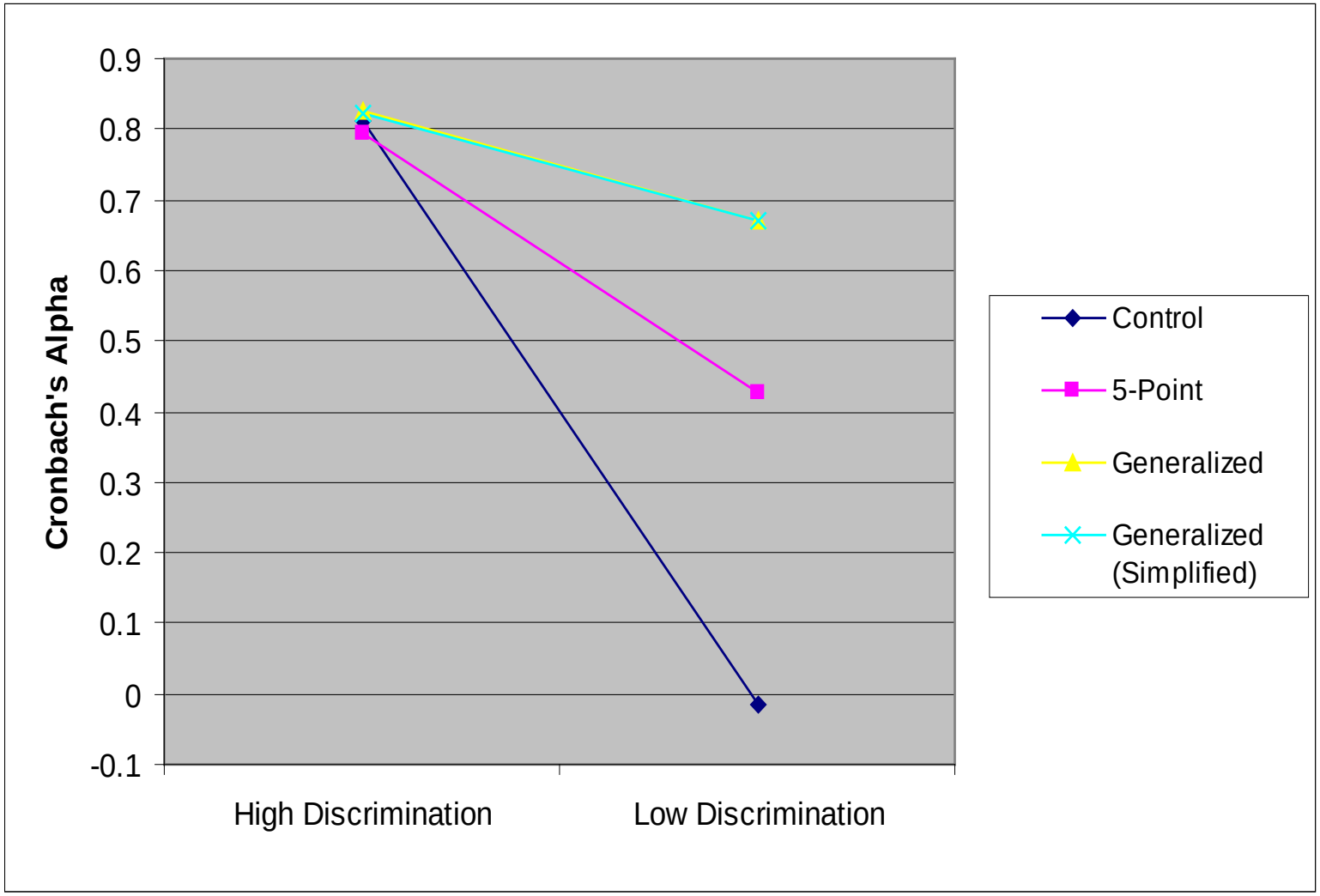


Figure 20 – Reliability by Discrimination (Easier Items).

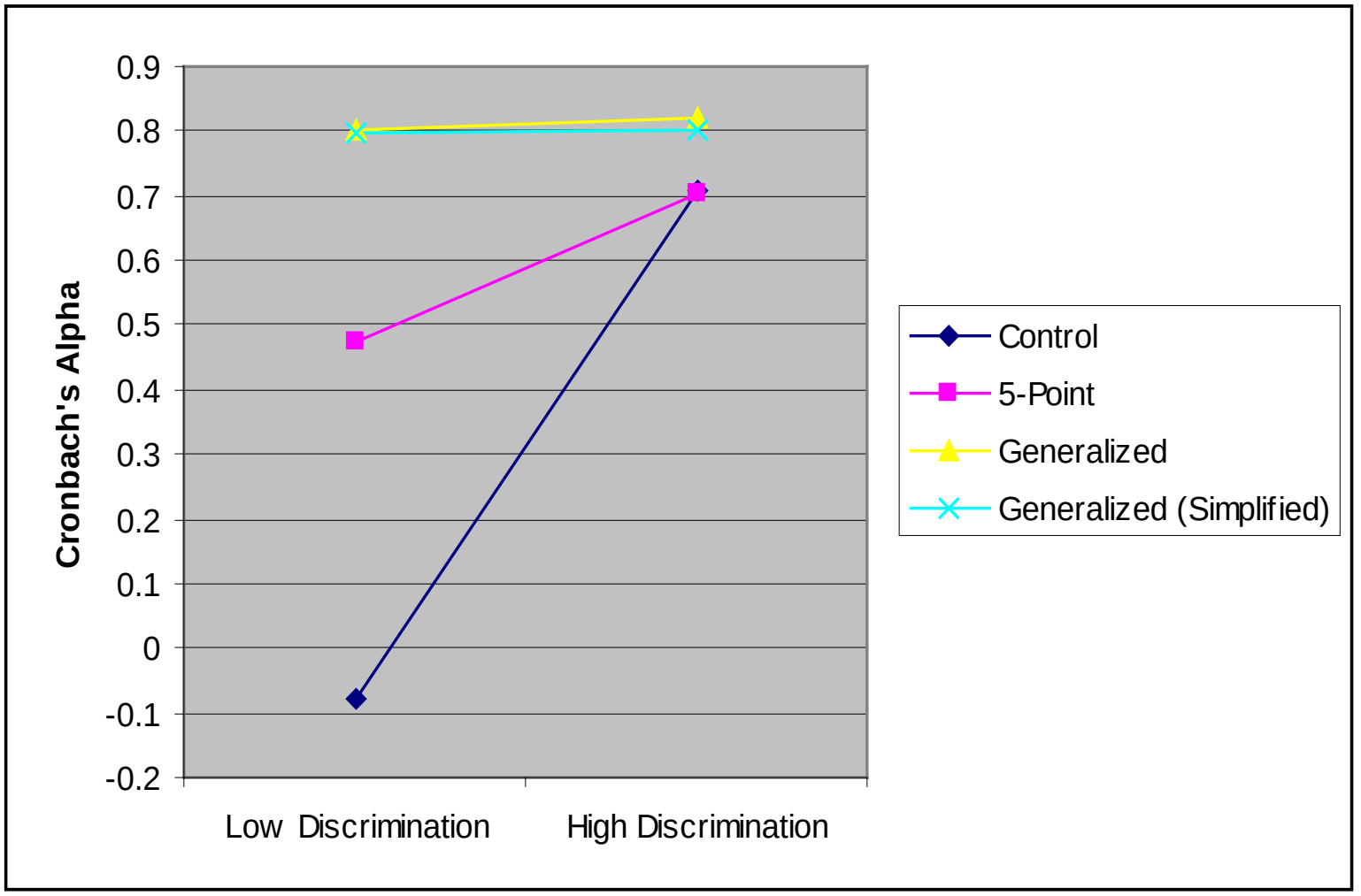


Figure 21 – Reliability by Discrimination (More Difficult Items).

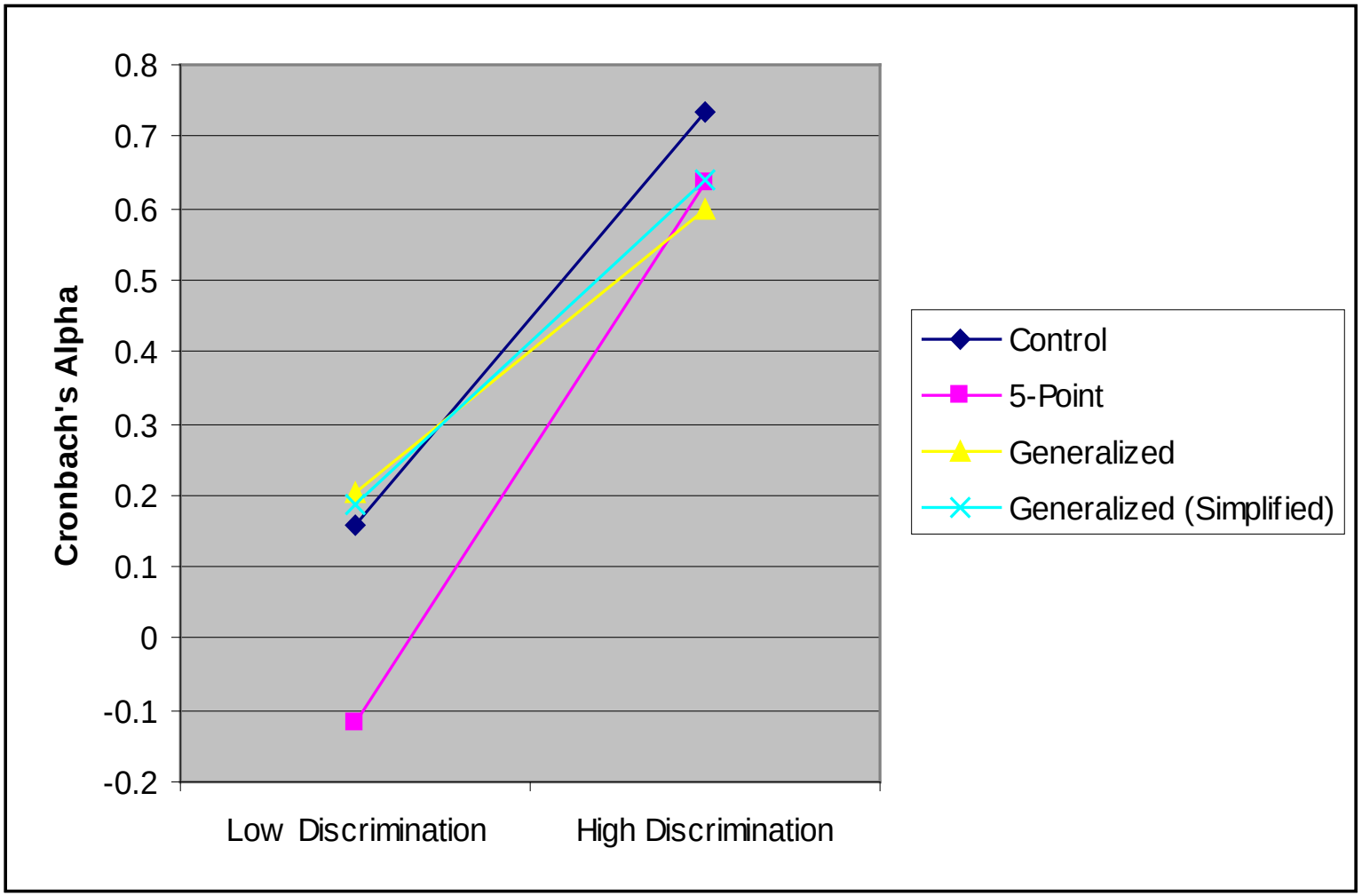


Figure 22 – Reliability by difficulty on only the 20 easiest items.

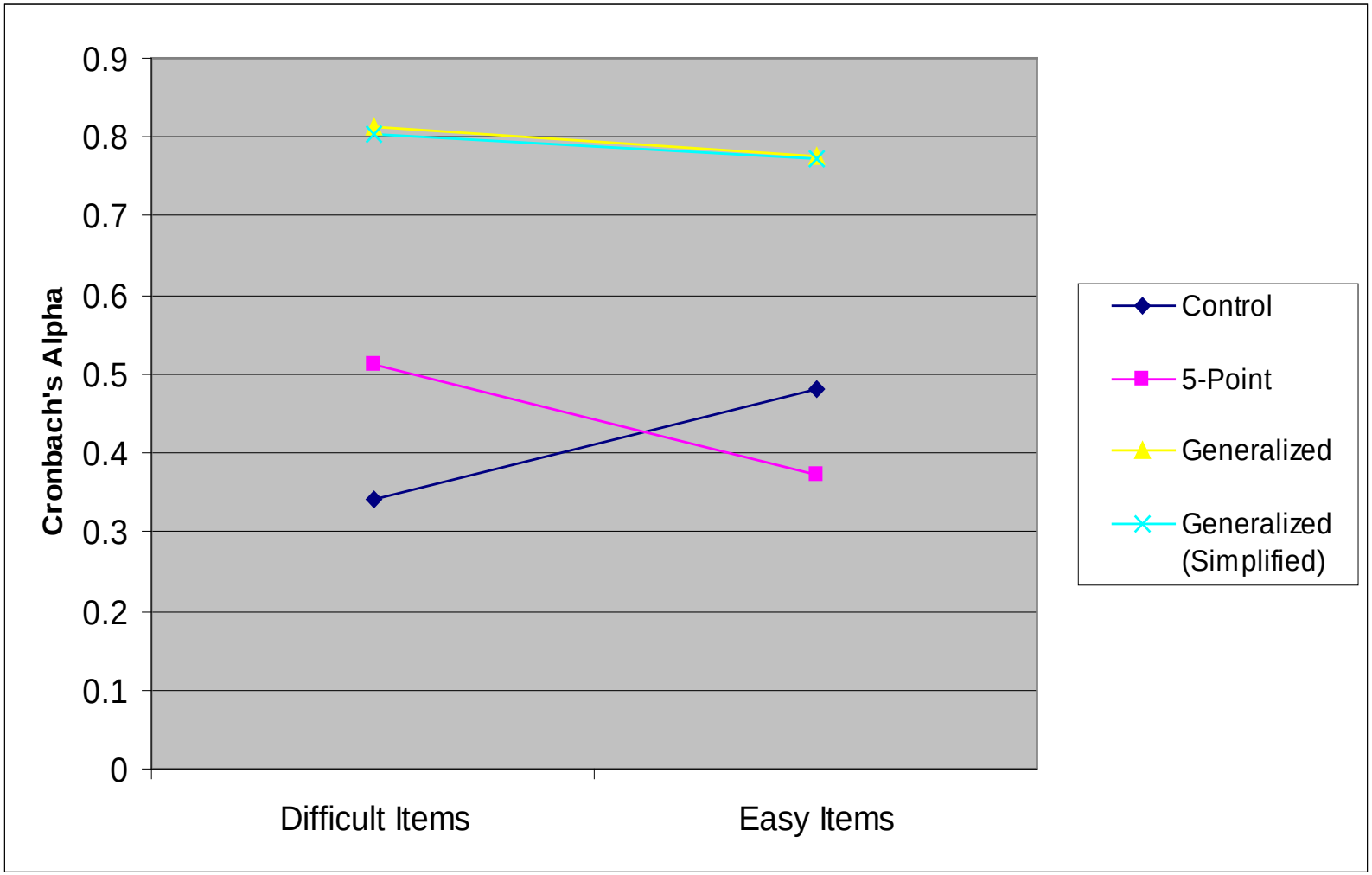


Table 1: A Table of Responses and Weights.

Response in A	Response for B	Response for C	Weight for A	Weight for B	Weight for C
Very sure			1	0	0
Fairly sure			0.66	0.16	0.16
	Very sure		0	1	0
	Fairly sure		0.16	0.66	0.16
		Very sure	0	0	1
		Fairly sure	0.16	0.16	0.66
Pure guess	Pure guess	Pure guess	0.33	0.33	0.33

Table 2: A Table of Responses and Integers.

Response in A	Response for B	Response for C	PA* to A	PA to B	PA to C
Very sure			6	0	0
Fairly sure			4	1	1
	Very sure		0	6	0
	Fairly sure		1	4	1
		Very sure	0	0	6
		Fairly sure	1	1	4
Pure guess	Pure guess	Pure guess	2	2	2

* - PA = points allocated

Table 3: Results of Pilot Data Simulation #1

	Correlation with True Score	R-Square Value	R-Square Change (from standard)	Test Reliability
Standard Scoring	.9453	.8937	n/a	.907
Raw probability	.9857	.9717	.0780	.998
Ten-point allocation	.9851	.9705	.0768	.997
Five-point allocation	.9871	.9744	.0807	.996
Three-point allocation	.9824	.9652	.0715	.993
Two-point allocation	.9812	.9629	.0692	.988

Table 4: Results of Pilot Data Simulation #2

	Correlation with True Score	R-Square Value	R-Square Change (from standard)	Test Reliability
Standard Scoring	.9453	.8937	n/a	.907
Five-point allocation	.9845	.9693	.0755	.994
Two-point allocation	.9770	.9545	.0608	.976

Table 5: Test reliability as a function of test length.

Number of items	5	10	15	20	25	30	35	40	45	50	55	60
5-point allocation	.944	.967	.978	.984	.987	.988	.990	.991	.992	.993	.994	.994
2-point allocation	.802	.891	.917	.939	.948	.953	.960	.964	.967	.971	.974	.976
Traditional test	.510	.696	.754	.797	.823	.838	.856	.868	.877	.888	.900	.907

Green shows reliability above .90, yellow above .80, blue above .70, and red below .70.

NOTE: For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Table 6: Effect of test difficulty on correlation with 'True Score' for 60-item test.

	Correlation with True Score	R-Square Value	R-Square Change (from traditional)
Normal Test			
Five-point allocation	.9845	.9693	.0755
Two-point allocation	.9770	.9545	.0608
Easy Test			
Five-point allocation	.9657	.9326	.0793
Two-point allocation	.9577	.9172	.0639
Difficult Test			
Five-point allocation	.9104	.8289	.1759
Two-point allocation	.8931	.7977	.1447

Table 7: Effect of test discrimination on correlation with 'True Score' for 60-item test.

	Correlation with True Score	R-Square Value	R-Square Change (from traditional)
Normal Test			
Five-point allocation	.9845	.9693	.0755
Two-point allocation	.9770	.9545	.0608
Low Discrim. Test			
Five-point allocation	.9895	.9792	.1125
Two-point allocation	.9759	.9524	.0857
High Discrim. Test			
Five-point allocation	.9803	.9610	.0562
Two-point allocation	.9751	.9509	.0462

Table 8 – Reliability and Validity Across Conditions for 60 Item Test.

	Reliability	Validity	
Test Form	Cronbach's Alpha	College GPA	ACT
Control	0.761	0.237	0.710*
5-Point Confidence	0.803	0.385*	0.571*
Generalized Con.	0.862	0.295*	0.631*
Generalized Con. (Simplified)	0.862	0.299*	0.628*

Correlations significant at the $p < .05$ level starred.

Table 9 – Reliability by Test Length and Condition.

# of Items/ Condition	5	10	15	20	25	30	35	40	45	50	55	60
Control	0.211	0.371	0.387	0.462	0.445	0.575	0.616	0.672	0.681	0.732	0.729	0.761
5-Point	0.405	0.565	0.503	0.614	0.632	0.700	0.704	0.747	0.767	0.790	0.789	0.803
Generalized	0.360	0.645	0.614	0.730	0.759	0.799	0.833	0.848	0.870	0.866	0.862	0.862
Generalized (Simplified)	0.308	0.624	0.588	0.713	0.746	0.792	0.828	0.843	0.863	0.866	0.863	0.862

Green shows reliability above .80, yellow above .70, blue above .60, and red below .60.

Table 10 – Correlations of Individual Differences with CT Use (5-point condition).

		Use of Confidence Testing
Test-Specific Self Efficacy	Pearson Correlation	-.126
	Sig. (2-tailed)	.335
	N	61
General Anxiety	Pearson Correlation	-.203
	Sig. (2-tailed)	.117
	N	61
Test Anxiety	Pearson Correlation	.149
	Sig. (2-tailed)	.250
	N	61
Risk Taking	Pearson Correlation	-.028
	Sig. (2-tailed)	.828
	N	61
Cautiousness	Pearson Correlation	-.098
	Sig. (2-tailed)	.454
	N	61

Table 11 - Correlations of Individual Differences with CT Use (Generalized condition).

		Use of Confidence Testing
Test Specific Self Efficacy	Pearson Correlation	-.167
	Sig. (2-tailed)	.211
	N	58
General Anxiety	Pearson Correlation	-.039
	Sig. (2-tailed)	.772
	N	58
Test Anxiety	Pearson Correlation	.248
	Sig. (2-tailed)	.061
	N	58
Risk Taking	Pearson Correlation	.170
	Sig. (2-tailed)	.202
	N	58
Cautiousness	Pearson Correlation	-.132
	Sig. (2-tailed)	.324
	N	58

Table 12 - Correlations of Individual Differences with CT Use, Controlling for Overall Score (5-Point Condition).

		Use of Confidence Testing
Test Specific Self Efficacy	Pearson Correlation	.024
	Sig. (2-tailed)	.854
	df	58
General Anxiety	Pearson Correlation	-.214
	Sig. (2-tailed)	.101
	df	58
Test Anxiety	Pearson Correlation	.136
	Sig. (2-tailed)	.301
	df	58
Risk Taking	Pearson Correlation	-.112
	Sig. (2-tailed)	.396
	df	58
Cautiousness	Pearson Correlation	.064
	Sig. (2-tailed)	.626
	df	58

Table 13 - Correlations of Individual Differences with CT Use, Controlling for Overall Score (Generalized Condition).

		Use of Confidence Testing
Test Specific Self Efficacy	Pearson Correlation	-.138
	Sig. (2-tailed)	.305
	df	55
General Anxiety	Pearson Correlation	-.037
	Sig. (2-tailed)	.784
	df	55
Test Anxiety	Pearson Correlation	.100
	Sig. (2-tailed)	.459
	df	55
Risk Taking	Pearson Correlation	.026
	Sig. (2-tailed)	.847
	df	55
Cautiousness	Pearson Correlation	-.056
	Sig. (2-tailed)	.681
	df	55

Table 14 - Correlations of Individual Differences with CT Use, Controlling for Overall Score (Both Confidence Conditions).

		Use of Confidence Testing
Test Specific Self Efficacy	Pearson Correlation	-.054
	Sig. (2-tailed)	.564
	df	116
General Anxiety	Pearson Correlation	-.076
	Sig. (2-tailed)	.411
	df	116
Test Anxiety	Pearson Correlation	.008
	Sig. (2-tailed)	.928
	df	116
Risk Taking	Pearson Correlation	-.023
	Sig. (2-tailed)	.804
	df	116
Cautiousness	Pearson Correlation	.028
	Sig. (2-tailed)	.765
	df	116

REFERENCES

REFERENCES

Adams, J.K. (1961). Realism of Confidence Judgments. *Psychological Review,* 68, 33-45.

Baranski, J.V. & Petrusic, W.M. (1998). Probing the Locus of Confidence Judgments: Experiements on the Time to Determine Confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 929-945.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Coombs, C.H. (1953). On the use of objective examinations. *Educational and Psychological Measurement, 13*, 308-310.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology, 13*, 87-123.

Dressel, P.L & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13*, 574-595.

Ebel, R.L. (1965). Confidence Weighting and Test Reliability. *Journal of Educational Measurement, 2*, 49-57.

Ebel, R.L. (1968). Review of 'Valid Confidence Testing – Demonstration Kit'. *Journal of Educational Measurement, 5*, 353-354.

Echternacht, G.J. (1972). The use of confidence testing in objective tests. *Review of Educational Research, 42*, 217-236.

Erev, I., Wallsten, T.S., & Budescu, D.V. (1994). Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes. *Psychological Review, 101*, 519-527.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Hambleton, R.K., Roberts, D.M., & Traub, R.E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement, 7*, 75-82.

Hevner, K. (1932). A method of correcting for guessing in true-false tests and empirical evidence in support of it. *The Journal of Social Psychology, 3*, 359-359-362.

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (*In Press*). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology.*

Jacobs, S.S. (1971). Correlates of unwarranted confidence in responses to objective test items. *Journal of Educational Measurement, 8,* 15-20.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39,* 103-129.

Koehler, R.A. (1971). A comparison of the validities of conventional choice testing and various confidence marking procedures. *Journal of Educational Measurement, 8,* 297-303.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for Confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107-118.

Krueger, N. & Dickson, P.R. (1994). How believing in ourselves increases risk taking: Perceived Self-Efficacy and Opportunity Recognition. *Decision Sciences, 25,* 385-400.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* New Jersey, Lawrence Erlbaum Associates.

McKenzie, C.R.M., Wixted, J.T., Noelle, D.C., & Gyurjyan, G. (2001). Relation Between Confidence in Yes-No and Forced-Choice Tasks. *Journal of Experimental Psychology: General, 130,* 140-155.

Michael, J.J. (1968). The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement, 5,* 307-314.

Pleskac, T.J. & Busemeyer, J.R. (2010). Two-Stage Dynamic Signal Detection: A Theory of Choice, Decision Time, and Confidence. *Psychological Review, 117,* 864-901.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danmarks Paedagogiske Institut, Copenhagen.

Reckase, M.D. (2009). *Multidimensional Item Response Theory.* New York, Springer.

Wainer, H. (2000). *Computerized Adaptive Testing: A Primer.* New Jersey, Lawrence Erlbaum Associates.

Roskam, E.E. (1987). Toward a psychometric theory of intelligence. In E.E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology.* Amsterdam: North-Holland.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report 79-4). Knoxville: University of Tennessee (Department of Psychology).

Sarason, I.G. & Stoops, R. (1978). Test anxiety and the passage of time. *Journal of Consulting and Clinical Psychology, 46*, 102-109.

Scheiblechner, H. (1985) Psychometric models for speed-test construction: The linear exponential model. In S.E. Embretson (Ed.), *Test design: Developments in psychology and education*. New York, Academic Press.

Selig, E.R. (1972). Confidence testing comes of age. *Training and Development Journal, 26*, 18-22.

Shuford, E.H. & Massengill, H.E. (1967). *Valid Confidence Testing – Demonstration Kit.* Boston, The Shuford-Massengill Corporation.

Soderquist, H.O. (1936). A new method of weighting scores in a true-false test. The *Journal of Educational Research, 30*, 290-292.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519

van der Linden, W.J. (2006). A Lognormal Model for Response Times on Test Items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Linden, W.J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5-20.

van der Linden, W.J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement, 47*, 92-114.

van der Linden, W.J., Entink, R.H.K., & Fox, J.P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*, 327-347.