The Binormal Hypothesis of Specific Learning Disabilities

STEPHEN ANTHONY ALBONE

The Binormal Hypothesis of Specific Learning Disabilities

Stephen Anthony Albone

The concept of specific learning disabilities has its roots in the medical literature of the nineteenth century. According to the medical model the cause of specific learning disabilities are presumed to lie in specific cognitive dysfunctions. This hypothesis predicts two qualitatively distinct types of learner and a bimodal distribution of assessment scores. Evidence for bimodality has been sought in the distribution of residuals generated from the regression of standardised measures of attainment on IQ, however this technique has been widely criticised. Recent advances in computer adaptive assessment, coupled with Rasch interval level measurement, have opened up the possibility of seeking evidence for bimodality in the distribution of assessment scores directly.

In the present study the binormal distribution was developed as a model for describing bimodality. The binormal distribution is conceived as two superimposed normal distributions and is defined by five parameters. The algebraic relationship between the five parameters was first determined, and then a methodology was developed for deriving objective estimates of those parameters. The methodology was applied to a unique dataset of over 80,000 children aged between seven and eleven years of age, and across four assessment domains; picture vocabulary, reading, mathematics and arithmetic.

The methodology was found to be sensitive to factors that might influence the shape of the distribution of assessment scores such as gender, number of years of schooling, and ceiling effects, and this affected its utility. Nevertheless evidence was found for the existence two qualitatively distinct groups of reader. The pattern in these results was consistent with a developmental transition from beginning to fluent reader. Evidence was also found for a developmental lag between boys and girls, which would explain the higher prevalence of dyslexia reported for boys in many studies. The methodology produced inconsistent results when applied to the other assessments, and no evidence was found to either confirm or disprove the existence of specific dysfunctions as predicted by the medical model.

The Binormal Hypothesis of Specific Learning Disabilities

Stephen Anthony Albone

Submitted for the degree of
Doctor of Philosophy

Department of Education
Durham University
2010

Contents

List of Tables

List of Figures

I would like to thank the Centre for Evaluation and Monitoring at Durham University for their sponsorship of this research. Particular thanks go to Brian Henderson and Matthew Durham for their assistance in collating the raw data files, and Paul Jones for advice concerning chapter 2. Finally I would like to extend a heartfelt thank you to Peter Tymms and John Adams for their guidance, support and patience in what has been at times a difficult journey.

For Elizabeth

Chapter 1: The Background and Context of the Study

1. Introduction

The study of special educational needs goes back to the medical literature of the nineteenth century. However it was not until much later that researchers from non-medical disciplines began to take an interest in this area. With that interest came much debate concerning the nature of special educational needs and the terms used to define it. This remains one of the most contentious issues in psychological and educational research.

A central theme of the debate concerns whether or not there are qualitatively distinct subpopulations of learners. Presented here is a historical summary of that debate insofar as it pertains to that specific issue. This leads on to the rationale for the present study.

2. Early Case Studies of Reading Disabilities

The origin of the concept of learning disabilities has its roots in the medical literature of the nineteenth century. Early case studies of patients with reading disorders were to introduce a number of terms and ideas that would shape our present understanding of the condition. Some of the key papers are outlined below.

In an 1887 monograph the German ophthalmologist, Professor Dr. Rudolph Berlin described six case studies of adult patients that showed particular difficulties with reading (Berlin, 1887, Wagner, 1973). Berlin had written of the same condition in a previous work, and had introduced the term dyslexia to describe it (Berlin, 1884). Following post-mortem examination of his patients Berlin attributed the probable cause of dyslexia to a brain dysfunction.

In the United Kingdom similar case studies were also reported, crucially however this literature included numerous examples of children that in spite of receiving every advantage lacked the specific capacity to learn to read

(Hinshelwood, 1900, Kerr, 1897, Morgan, 1896, Nettleship, 1901). In the absence of any obvious illness or injury it was concluded that these children probably suffered from a congenital condition known as word blindness. It was Hinshelwood that first made the distinction between this congenital form and the acquired word blindness that was observed in adults (Hinshelwood, 1896).

The first report of children having word blindness in Germany was made by Foerster (Foerster, 1905). In doing so he made a distinction between word blindness that was associated with mental retardation, and that which presented as a specific cognitive deficit. He also suggested that the two forms of word blindness had a different underlying neurological cause.

Thus at an early date the hypothesis for the existence of two distinct causes of reading failure was established. On the one hand the inability to learn to read might be due to general cognitive deficits affecting the intelligence of the individual. Alternatively the condition might be characterised by particular cognitive impairments, and presumed to be due to a specific dysfunction of the central nervous system. This hypothesis has become known as the medical model of learning disabilities. An extensive account of the pioneering work in this field is provided by Anderson and Meier-Hedde (Anderson and Meier-Hedde, 2001).

3. The Scholastic Disabilities Model

In the United Kingdom the concept of learning disabilities was further developed by Schonell, but from an educational perspective (Schonell, 1935). Schonell was interested in comparing measures of the intelligence quotient (IQ) with a related measure of scholastic achievement known as the accomplishment quotient (AQ). Comparison of these two measures enabled Schonell to identify children that were underachieving with respect to their IQ. He went on to define three classes of underachievement, which he termed scholastic disability.

1. *The innately dull who are backward because of inferior intellectual powers and hence need teaching differing qualitatively and quantitatively from that accorded to normal children.*
2. *Those of unimpaired general intellectual powers who are backward in school work and who simply need continuous schooling, individual assistance or special coaching to overcome this handicap.*
3. *The supernormal pupils whose disparity between IQ and AQ should hardly be considered as backwardness, a more suitable term being retardation; their backwardness is more apparent than real. Such scholars require increased private study and a fuller curriculum to extend them to the limit of their intellectual capacities.*

Whilst there is no explicit reference to the medical model in Schonell's definition it is nevertheless implicit that there are different underlying causes for academic failure. It is the recognition of these different causes that is vital because it informs the choice of remediation strategy for each child.

Schonell was principally interested in the profile of scores across different academic disciplines, which prompted a further complication to his model. Schonell observed that children may show scholastic disability in some areas but not others. He defined these as specific disabilities.

Schonell's model introduced two important concepts into the field of learning disabilities. The first was the notion of a discrepancy between academic potential and achievement. The second was the recognition of different types of learning failure requiring qualitatively different remediation strategies. However there are difficulties that arise as a consequence of its dependence on cut-scores to define intelligence and accomplishment as either low, normal or high. This presents certain practical as well as philosophical difficulties when applying the model. A second weakness in the model stems from the practice of defining underachievement as the arithmetical difference in IQ and AQ measures, a practice that fails to take account of systematic patterns in the magnitude of measurement error in the data. A final difficulty with using the model results from its emphasis on underachievement without a fuller consideration of low achievement.

4. The Regression Model of Underachievement

As mentioned in the previous section one difficulty with Schonell's scholastic disabilities model is that the practice of defining underachievement in terms of an arithmetical difference between IQ and AQ measures fails to properly account for the measurement error in the data. According to classical test theory a person's observed assessment score consists of two components, the true score and the measurement error (Hopkins, 1998). Thus the assessment score is only an estimate of the true ability. The magnitude of the error component varies between individuals for a variety of reasons. A difficulty arises in the interpretation of assessment scores because the distribution of measurement error is not random. The value of the error may be positive, resulting in an assessment score that overestimates the true score, or it may be negative resulting in an underestimate of the true score. There is a systematic tendency for individuals with low scores to have a large negative error, whilst high scores tend to be associated with large positive errors. This bias in the data can be controlled for using the regression methodology.

The regression methodology can be applied when a causal effect between two variables is justified on theoretical grounds. In the case of Schonell's model it is assumed that AQ is somehow dependent on IQ, which is regarded as independent. To apply the regression method the dependent variable (AQ) is plotted against the independent variable (IQ). A line of best fit is then applied to the data in such a way that the sum of squares of the vertical distances between each data point and the line of best fit, known as the regression line, is minimised. The vertical distance from each data point to the regression line is called the residual because it represents that part of the AQ measure that is left over after the IQ of that person has been taken into account. This effectively controls for the systematic distribution of measurement error in the dependent variable (AQ). The residual thus generated provides a measure of the under- or over-achievement after the bias in measurement error has been corrected for. A full description and justification of the regression approach is given by Thorndike (Thorndike, 1963).

5. Application of the Regression Model: The Isle of Wight Studies

The regression methodology was applied by William Yule, Michael Rutter and colleagues in a series of influential papers (Rutter, 1978, Rutter and Yule, 1975). In those studies the researchers gathered reading and IQ measures from all of the children in three year groups located in the Isle of Wight (aged 9 to 11 years). Follow up data were gathered on two of these cohorts at age 14. A fifth population of children from an inner London borough was also assessed. The number of individuals constituting each of the five study populations ranged from 1134 to 2113. The data so gathered were used to define two classes of poor readers. The first group consisted of those individuals with reading scores that were at least 2 years 4 months below the average expected for their age. Such children were classified as low achievers and described as having a general reading backwardness. The researchers then used a multiple regression approach to produce a predicted level of reading for each child after taking into account their chronological age and IQ. Children with reading scores that fell at least 2 years 4 months below that predicted by the regression model were classified as underachievers and described as having a specific reading retardation.

A difficulty with this classification approach was that a high proportion of weak readers fell into both categories. However this was justified as an important way forwards on two grounds. The first concerned the differential educational prognosis of backward readers compared to retarded readers and the second was concerned with the shape of the distribution of residual scores in the population as a whole. Each of these lines of evidence is described below.

In the follow up study of the two youngest Isle of Wight cohorts, Yule found evidence of a different educational prognosis for the two classifications of poor reader (Yule, 1973). Initially the two groups had a similar average reading age, about 33 months below the cohort average. However they differed in their average IQ scores. The reading retarded (underachieving) children had an average IQ score similar to that of the control population, whilst that of the reading backward (low achieving) group was about one standard deviation

lower. The children were given follow up assessments of reading accuracy, spelling and arithmetic. A striking aspect of the results was that the retarded readers made significantly better progress than the backward readers in arithmetic, but in reading and spelling the opposite result was found. It was concluded that the distinction between the two groups of poor reader had a valid educational significance. Later studies also found evidence for differences between the two groups of poor readers in terms of sex ratio, incidence of neurological disorder and pattern of neuro-developmental deficits (Rutter and Yule, 1975).

According to the medical model there are two qualitatively distinct subpopulations of reader which might be referred to as dyslexics and typical readers. As distinct subpopulations they would each be expected to have a quite different distribution of reading assessment scores, with the dyslexic subpopulation having a lower mean score. The medical model therefore predicts a bimodal distribution of scores in the population. Given that it is entirely possible that the distribution of scores in the two subpopulations might have considerable overlap, with the most able dyslexic children gaining higher scores than the weakest of the typical readers, this bimodality would not necessarily appear as two distinct peaks in the distribution of reading scores (Everitt, 1981, Fleiss, 1972). In fact it was asserted by Critchley that the distribution of reading scores has a hump in the left-hand tail due to the presence of dyslexics in that part of the reading abilities range (Critchley, 1970, Yule et al., 1974).

Having found evidence for two educationally distinct groups of poor readers Yule and colleagues sought and found empirical evidence to support Critchley's assertion (Yule et al., 1974). In their study, the distribution of residuals obtained from the regression of standardised reading scores on standardised non verbal intelligence scores for five populations of children were analysed. It was shown that there were significantly more children in the left hand tail of the distributions than would be expected by chance if the residuals were normally distributed in each population. The authors acknowledged that according to Gruenberg (Gruenberg, 1966) such humps can emerge as an artefact when the abilities of the sample tested are broader than those on which the assessment was

standardised, but argued that this was not an issue in this instance. They concluded that the presence of the hump provided evidence for the existence two distinct groups of weak readers.

6. Specific Criticisms of the Isle of Wight Studies

Reactions to the findings of the Isle of Wight studies have been mixed. Silva and colleagues presented the results from a study of 952 nine-year-olds from New Zealand that were in broad agreement with respect to the differences between the reading backward and reading retarded groups (Silva et al., 1985). The authors claimed that the differences they observed may be of aetiological significance, although Rutter and Yule had previously specifically stated that their findings did not support that view, despite the apparent similarities between their retarded readers and many of the characteristics associated with dyslexia. Specific differences in early literacy and phonological processing skills were also found in a study of 453 Australian children over their first three years in school (Jorm et al., 1986). However in a re-evaluation of the New Zealand study the validity of the apparent differences between the retarded and backward readers was questioned (Share et al., 1987). The objection to the previous conclusions was based upon the assumed relationship between the reading and IQ measures, which the authors argued was really correlative rather than causative in nature. It was concluded that there was in fact no evidence for bimodality, and that it was better therefore to treat under-achievement as a continuum.

Another line of criticism challenged the underlying assumption that the residual scores of reading on IQ are normally distributed (van der Wissel and Zegers, 1985). The authors argued that this would only be expected if three conditions were met:

1. *The reading scores are normally distributed.*
2. *The regression of reading scores on IQ is linear.*
3. *The variance of the conditional reading score distribution given the IQ level is constant over all IQ levels (homoscedasticity).*

They concluded that the apparent hump observed in the Isle of Wight data might simply result from a ceiling on the reading test.

Attempts to replicate the findings have also met with mixed results. On the one hand Stevenson found evidence for a hump in the residual scores of eleven-year-olds for reading and spelling, but not mathematics (Stevenson, 1988). Rodgers however failed to reproduce the findings using a reading assessment that had no ceiling (Rodgers, 1983). He gave four difficulties with the original study that might have affected the outcomes of the analysis. These were as follows:

1. *The reading tests employed had an acknowledged ceiling, with a resultant deviation from linearity in the regression functions employed.*
2. *The histograms of residual scores that were presented might more properly be regarded as negatively skewed, rather than bimodal.*
3. *The prevalence estimates for specific reading retardation varied considerably between different groups according to which tests were used.*
4. *The test score distributions may have deviated from normality in such a way as to affect the shape of the distribution of residual scores.*

In essence Rodgers was arguing that the observed distribution of residual scores was simply an artefact resulting from limitations in the assessments employed. Shaywitz and colleagues also failed to find any evidence for a hump in the distribution of reading residual scores (Shaywitz et al., 1992b). They also reported that only 28% of children classified as under-achieving in Grade 1 received the same classification two years later. The lack of stability in reading disability classification is unexpected if the model were valid, and the authors concluded that reading disabilities fall on a continuum with typical reading acquisition that can be modelled using the normal distribution.

7. The Rejection of the IQ-Discrepancy Model

Although there are earlier discussions in the literature the first strong rejection of the IQ-discrepancy model of learning disabilities came from Siegel (Siegel, 1989). That rejection was largely based on arguments concerning the validity of various assumptions that underpin the model, but she also rejected empirical evidence for the cognitive differences between individuals identified as low

achievers compared to under achievers. Although her objections were specifically discussed in relation to reading disabilities she stated that her arguments could be generalised to other forms of academic achievement. She went on to gather further empirical evidence to support her view (Siegel, 1992). A study by Shaywitz and colleagues also concluded that there are more similarities than differences between the two groups of poor readers (Shaywitz et al., 1992a).

The model was also rejected by Aaron following a review of the evidence for the predicted qualitative differences between the subtypes of weak reader (Aaron, 1997). As well as the validity of the model Aaron also questioned its utility in informing different remediation strategies for identified individuals. As an alternative Aaron proposed the reading component model as a more appropriate approach. According to this model all poor readers would be assessed for specific weaknesses in particular reading skills such as word decoding and comprehension and given remediation to specifically address any weaknesses that were found. In a later study Pennington and colleagues did find evidence for a differential effect in component reading processes between the two reading disability subtypes (Pennington et al., 1992). However the same study found no evidence for an underlying genetic cause for the differences, an observation that would have supported the medical model. In a quite different approach Vellutino and colleagues demonstrated that IQ-discrepancy alone failed to distinguish between two important classes of poor reader, specifically those that did and did not respond to remediation (Vellutino et al., 2000).

Criticism of the IQ-discrepancy model is not restricted to reading disabilities. It has also been shown in a study addressing arithmetical learning disabilities that there is little evidence for a difference between low-achieving and under-achieving arithmeticians with respect to performance on working memory tasks and arithmetical word problems (Jiménez González and Garcia Espínel, 1999).

Perhaps the most vehement critique of the IQ discrepancy model has been made by Stanovich (Stanovich, 2005). As well as reiterating the points made by Siegel and Aaron, he argues that a fundamental difficulty with the IQ-discrepancy

model is its dependence on a causative unidirectional relationship between IQ and attainment and points to evidence that the relationship is in fact reciprocal, at least in the case of reading (Stanovich, 1991, Stanovich, 1993).

In defence of the discrepancy model Kavale has argued that it remains useful if it is simply regarded as an operational definition of underachievement (Kavale, 2001). In other words it is a practical tool to inform identification, which is a considerably weaker role than its use for defining learning disabilities. In a meta-analysis of 46 studies Stuebing and colleagues found little evidence to support the validity of the IQ-discrepancy classification (Stuebing et al., 2002). Nevertheless they argued that it did not necessarily mean that the concept of learning disabilities is invalid, rather that the operational implementation of the concept is flawed. It is a problem with all studies that utilise the regression model that there is a considerable overlap between low achieving and under achieving individuals.

Dissatisfaction with the discrepancy model has prompted Vaughn and Fuchs to propose an alternative framework for defining learning disabilities (Vaughn and Fuchs, 2003). The response-to-intervention model shifts the emphasis from student deficits to student outcomes, however the diagnostic validity of this approach has also been questioned by Kavale who councils against the wholesale rejection of the discrepancy model (Kavale, 2005). It seems that in spite of the practical and philosophical obstacles, the discrepancy model of learning disabilities has an appeal that is difficult to overcome. That appeal may lie in the face validity of the underlying concept of qualitatively distinct subtypes of learners, and perhaps what is really needed is a different toolkit to explore that idea.

8. The Definition of Specific Learning Disabilities

The term "learning disability" was introduced by Samuel Kirk in 1962 (Kirk, 1962). According to Kirk:

> *A learning disability refers to a retardation, disorder, or delayed development in one or more of the processes of speech, language, reading, writing, arithmetic, or other school subject resulting from a psychological handicap caused by a possible cerebral dysfunction and/or emotional or behavioral disturbances. It is not the result of mental retardation, sensory deprivation, or cultural and instructional factors.*

A central feature of this and many later definitions of learning disabilities is the controversial notion that the presumed cause of the condition is a dysfunction of the central nervous system. It is an idea that is derived directly from the medical literature and one that has profound implications. If the presumption is correct, it means that the population consists of qualitatively distinct subpopulations of learners. If not it means that learners form a single continuum of ability, and that the differences between the weak and able are merely quantitative. The truth or otherwise of the presumption has important practical implications for how we identify and remediate learning disabilities, as well as philosophical implications for how we view affected individuals.

Another important feature of Kirk's description of learning disabilities that is often repeated in other definitions is the exclusion clauses that form the final sentence. Such clauses also have their origin in the medical literature and reinforce the notion of qualitatively distinct subpopulations of learners. The mental retardation clause in particular has been hugely influential in the adoption of operational definitions of learning disabilities that utilise regression discrepancy models. The use of such clauses have been criticised for defining learning disabilities in terms of what they are not, rather than what they are, and attempts have been made to remove both the exclusion clauses and aetiological component from the definition of learning disabilities (Wepman, 1975).

According to Kavale a third important feature of Kirk's definition is that it introduced the notion of intra-individual differences (Kavale, 2001), although

this was also hinted at in the earlier work of Schonell (Schonell, 1935). The idea that a learning disability may differentially affect particular areas of academic attainment, and be independent of normal functioning in other cognitive domains, has prompted the introduction of the term *'specific learning disabilities'* into the literature. Interestingly this has led to the use of the unqualified term *'learning disabilities'* to refer to conditions that result from mental retardation, which is in complete contradiction to the exclusion clause in Kirk's original definition.

This contradiction in terminology illustrates a fundamental issue at the heart of definitions of learning disabilities. Kavelle and Forness argue that it is an inherent problem that interested parties simply define learning disabilities according to what they think it should be (Kavale and Forness, 2000). They state that:

> *Learning disabilities definitions, although useful, remain equivocal with respect to validity because they properly belong to the stipulative class of definition.*

They go on to state that:

> *In reality, stipulative definitions are only of generic usefulness and require transformation to be applied in practice. The most usual transformation is the operational definition, rules stipulating how the term is to apply to a particular case if specified operations yield certain characteristic results.*

According to Kavale and Forness operational definitions suffer from a number of philosophical and practical difficulties. The outcomes of operational procedures are sensitive to the reliability and validity of the operational indicators chosen to make the necessary observations. Validity is also an issue when it comes to matching an operational procedure with the underlying theoretical construct. It is possible to operarationalise anything but it doesn't follow that it has any meaning in reality. There is a real danger of simply defining learning disabilities in terms of what can easily be measured, but unless this is coupled to a valid theoretical model then our understanding of the condition is not enhanced. Ultimately the diagnosis and remediation of learning disabilities is dependent on

this understanding. They conclude that as the definition of learning disabilities has developed from one that is conceptually physiological to one that is behaviourally centred it has lost the power to inform us what learning disabilities actually are.

Perhaps the real issue behind this apparent paradox is an unrealistic expectation of what definitions of learning disabilities actually are. The situation is clarified by Snowling for whom the term specific learning difficulties refers to a *'statistical definition'* which *'carries no implication about the nature or aetiology'* of the condition (Snowling, 2005). For Snowling then specific learning disabilities provide an operational definition which may, or may not indicate a dysfunction, and can at best therefore offer the clinician a starting point for their diagnosis.

9. Rationale for the Present Study

My interest in this area stems from my role as a Research Associate at the Centre for Educational Management (CEM) based at Durham University. Part of my role involves the training of teachers in the interpretation of feedback generated by CEM's primary school level monitoring projects. It was in this role that I observed that one of the uses to which the data were put was to provide objective evidence to support the identification of children with special educational needs. This evidence was important because access to Educational Psychologists for the purposes of making a formal diagnosis is both limited and has cost implications.

At the time the CEM assessments available for the five to eleven year age range were exclusively presented in a paper and pencil format. The assessment system, known as Performance Indicators in Primary Schools or simply PIPS, uses standardised assessments of curriculum dependent measures (reading and mathematics) that are regressed against standardised measures of developed abilities (picture vocabulary and non-verbal ability) to provide an indication of over- or under-achievement. A fuller description of the assessments have been provided by Tymms and Albone (Tymms and Albone, 2002).

I was interested in how appropriate it was to use the PIPS data as a screening device for special educational needs, and had two principal concerns. The first of these was the accuracy of standardised assessment scores for atypically achieving children. Paper based standardised assessments are designed to discriminate between typically achieving children. As a child's score on such an assessment departs from the average the error of measurement on that score can increase enormously. This is not surprising if you consider the experience of a child that is academically very weak. Such a child may be able to tackle one or two of the easiest items in the assessment, but the rest of it may be unobtainable to them. If this is the case the child has effectively been presented with a two item assessment, irrespective of how many items the assessment actually contains. The best that we can say about the child is that they are very weak in the area that the assessment is designed to measure, but it not possible to state exactly how weak they are with any precision. The problem is just as acute for very able children for whom successful interaction with the majority of the assessment items is almost a given and who are only really tested by one or two of the most difficult items. We can conclude that such children are very able, but not exactly how able they are.

My second concern was prompted by the doubts that had been raised about the validity of the IQ-discrepancy model, particularly in the United States of America. There are some fundamental differences in the way that PIPS was being used as a screening tool in the UK, compared to the much stronger use of IQ-discrepancy as an operational definition of learning disabilities. In addition the developed abilities component of the PIPS assessment was never intended to represent a proxy measure for intelligence *per se*. Nevertheless some of the technical objections to the regression approach were worthy of reflection.

One way of approaching these concerns was to take advantage of the huge longitudinal dataset of PIPS assessment scores that had been collected over a number of years. It was reasoned that even if there were a high measurement error on a single assessment result for a child with special educational needs, that error would be reduced if the results from several assessments taken over several years were aggregated. This approach might also provide a way of looking at the

stability of scores over time. In addition, if enquiries were restricted to longitudinal patterns in curriculum dependent measures without reference to the developed ability measures it would completely side-step the issues concerning the application of the regression methodology.

In an early discussion along these lines Professor Peter Tymms proposed the hypothesis that if the population did indeed consist of qualitatively distinct subtypes of learner, then that might be revealed in the shape of the distribution of mean normalised scores. The rationale for this hypothesis is as follows.

One reason why previous studies have not looked for bimodality directly in measures of academic achievement is that the distribution of standardised test scores does not accurately reflect the true distribution of that ability in the population. According to classical test theory an individual's test score is made up of their 'true score' and the error on that score. However that error is not randomly distributed in the population. An assessment is made up of separate items. In a well constructed assessment the distribution of item difficulties will reflect the distribution of abilities in the population that the assessment is intended for. An assessment item can be thought of as being well targeted to a particular individual if the difficulty of that item is closely matched to their ability. If the item is too easy or too hard, that is if it is not well targeted, then it can reveal little about an individual's ability. From necessity standardised assessments are targeted towards typically achieving individuals, meaning that a relatively high proportion of items in the assessment will suit the average ability. As an individual's ability moves further from the average they will typically be presented with fewer and fewer well targeted items. Since the absolute magnitude of the measurement error is an inverse function of the number of well targeted items there is a systematic tendency for measurement error to increase as ability departs from the average. The matter is further complicated if the direction of the error is considered. If the error is positive it will cause the test score to be higher than the 'true score' that it is intended to estimate. If the error is negative it will cause the 'true score' to be underestimated. There is a systematic tendency for high scores to be associated with positive measurement error, whilst low scores are associated with negative measurement error. Since

the variation of measurement error with ability is largely the property of the assessment itself, and that error makes up a significant proportion of the overall score, it is unwise to infer much about the distribution of the 'true score' in the population. In short any evidence for bimodality that may be observed in the distribution of assessment scores may simply be an artefact of the assessment itself.

The process of standardising assessment scores effectively involves the linear transformation of raw scores to fix their mean and variance to some agreed value. The shape of the resulting distribution is exactly the same as that of the raw data. This manipulation may be taken a step further by forcing the data into a normal distribution. This process is called normalisation. When a group of children are assessed on two occasions and their mean normalised score determined it is found that the variance of those mean scores is less than that of the original normalised scores, although the mean remains the same. This is because the mean score will tend to have less measurement error than either of the individual assessment scores. If a third assessment is added to the analysis then the variance of mean normalised scores will shrink still further, and so on. Theoretically after an infinite number of assessments all of the measurement error will have been accounted for and the mean normalised score will be equivalent to the true score. Clearly it is not possible to conduct an infinite number of assessments; however it may be possible to gain some insight into the true-score distribution given a suitable longitudinal dataset. Specifically if the underlying distribution of ability is unimodal, then the distribution of mean normalised scores across several assessments will also tend to be unimodal. However, if the underlying distribution of ability is bimodal this will begin to reveal itself in the shape of the distribution of mean normalised scores. Furthermore, the shape of the distribution of mean normalised scores will reveal something of the nature of the relationship between the two sub populations. The medical model predicts that the smaller subpopulation will have the lower mean score. The resulting asymmetry in the distribution of mean normalised scores would be indicated by a negative skew. If the smaller subpopulation were to have the higher mean score a positive skew would be observed. If the two subpopulations were of equal size and variance then the distribution of mean normalised scores would remain

symmetrical, and the skew would be zero. However, even in this case evidence for bimodality would be revealed in an increased kurtosis in the distribution of mean normalised scores.

It was originally intended that the application of this idea would form the basis of the present thesis. Large amounts of longitudinal data were available and the process of deriving mean normalised scores was straightforward. Initial results were encouraging, suggesting evidence for bimodality in the distribution of reading scores, but not in mathematics. An attempt was also made to model the distribution of scores as the sum of two normal distributions (binormal distribution) and thereby derive parameter estimates for the distribution of each subtype of reader. The results of these analyses were presented at two conferences (Albone et al., 2006b, Albone et al., 2006a). However further development of the methodology revealed a number of problems. It was shown by use of randomised datasets having a binormal distribution that the normalisation process so confounded the scores that regression to the mean could never fully resolve the two subpopulations, thus making it impossible to derive accurate parameter estimates. The methodology for deriving parameter estimates was itself flawed, based as it was on the shape of the frequency histogram generated from the mean normalised scores. The parameter estimates generated in this way were found to be sensitive to the position of the category boundaries chosen for each bar of the histogram, so the methodology lacked robustness. The precision of the parameter estimates was dependent on the category width selected, and the selection of the midpoint of each category to represent its horizontal location introduced a systematic bias. In addition the process of aggregating the data into categories resulted in a second systematic bias according to the frequency of scores in each category. Since each bar of the histogram contributed a single data point to the model fit, scores that fell into infrequent categories had a disproportionate influence on the final result. Finally, and perhaps most significantly, it was found that if PIPS reading assessment scores were taken in pairwise combinations (there were six assessments and therefore fifteen pairs) the skewness observed in the mean of the standardised scores explained 71% of the variance observed in the skewness of the mean normalised scores. The equivalent figure with respect to kurtosis was found to be

35%. In other words the shape of the distribution of mean normalised scores was still heavily dependent on the distribution of raw assessment scores. Like the Isle of Wight studies before, the apparent bimodality might simply be an artifact of the assessments used.

These analyses prompted a review of the approach used and this line of investigation was abandoned. However this preliminary effort laid the groundwork for what is the subject of this thesis. The failure of the mean normalised scores approach ultimately resulted from the insufficient quality of the data used to evaluate it. A difficulty with standardised assessment results such as those used here is that the data generated is really at the ordinal level, but to gauge the true distribution of scores in the population requires interval level data. It has been shown that in the mid range standardised scores provide a good approximation of interval level data, but not in the tails of the distribution where children with learning disabilities are likely to reside (Cohen, 1979, Preece, 2002). A solution is to apply the Rasch measurement model to the data to obtain true equal interval measurement (Bond and Fox, 2001). However this is not sufficient on its own as it cannot compensate for the high measurement error in the distribution tails. That requires the use of a computer adaptive assessment. An assessment of this kind acts dynamically, selecting assessment items of appropriate difficulty according to each individual's preceding pattern of responses. This results in a high proportion of well targeted questions, maximizing the efficiency of the assessment whilst minimizing the measurement error, and in a way that is independent of the ability of the subject.

The aim of the present study was to determine if there is any evidence for bimodality in the distribution of assessment scores that would support the medical model of specific learning disabilities. This was done by analysing the scores of the entire population of four primary cohorts in Northern Ireland without reference to measures of ability or IQ. The data were gathered using a computerised adaptive assessment and processed using Rasch to obtain interval level measures. The central theme of the thesis is the development of the binormal distribution as a model of bimodality. The binormal distribution model is introduced, defined and described algebraically in chapter 2. In chapter 3 an

objective methodology is developed for determining binormal parameter estimates. Specific details of how that methodology was applied to the Northern Ireland data are given in chapter 4. Chapters 5 and 6 are concerned with a description of the results of that analysis. In chapter 7 the validity of these results are considered within the theoretical context of specific learning disabilities. In the final chapter the limitations of the methodology are discussed as are some of its potential applications.

10. Research Questions

The central theme of the present study is the application of the binormal distribution to the study of learning disabilities. The research questions that arise from this application are as follows:

1. Does the binormal distribution provide a suitable model for the investigation of bimodality in an epidemiological study of academic attainment in primary school children?
2. Is there any evidence for qualitatively distinct subtypes of learner in the population under study?
3. Is it possible to obtain valid and reliable parameter estimates for the distribution of assessment scores for different subtypes of learner within the population as a whole?
4. To what extent does the identification of distinct subtypes of learner support the medical model of specific learning disabilities? Is there any evidence for the existence of dysfunctions such as dyslexia and dyscalculia?
5. What are the implications of application of the binormal model to the identification of children with specific learning disabilities?

Chapter 2: The Binormal Distribution

1. Introduction

In this chapter the binormal distribution is introduced as a model for investigating data that has an underlying bimodal structure. The binormal distribution is conceived as the sum of two normal distributions. The model is then developed to determine the relationships between the parameters that define it.

2. Notation

The notation used to describe the binormal distribution has been adapted from the convention used in medical diagnostics (Pepe, 2003). Since the present study is primarily concerned with the distribution of scores on assessments of academic attainment the population is conceived as consisting of two qualitatively distinct subpopulations of learners. One of these subpopulations will tend to have low levels of attainment compared to the other. The other subpopulation may be regarded as having typical (or high) levels of attainment. The notation used is as follows:

D - attainment status (1 = low attainment, 0 = typical attainment)

$D, \overline{D}$ - subscripts for low and typically attaining subpopulations

ρ - prevalence of low attaining subpopulation $P(D=1)$

n_D - number of low attaining individuals

$n_{\overline{D}}$ - number of typically attaining individuals

N - total number of all individuals $= n_D + n_{\overline{D}}$

X - variate quantifying attainment

μ - population mean

$\overline{x}_D$ - mean of low attaining subpopulation

$\overline{x}_{\overline{D}}$ - mean of typically attaining subpopulation

σ^2 - population variance

s_D^2 - variance of low attaining subpopulation

$s_{\overline{D}}^2$ - variance of typically attaining subpopulation

3. The Normal Distribution

The normal distribution is a continuous probability distribution that is commonly used to model unimodal data. According to the central limit theorem the sum of a number of independent random variables with finite mean and variance tends to approach a normal distribution as the number of variables increases (Grinstead and Snell, 1997). This means that in real world situations where the magnitude of a particular measure is dependent upon a complex interplay of many underlying factors, the distribution of that measure over many observations will tend to have a normal distribution. This makes it possible to apply a relatively simple model to complex phenomena. One application of this kind of modelling is that it permits access to an array of powerful parametric statistical tools.

The probability density function (pdf) of the normal distribution in variate X is given by

$$P(X) = \frac{e^{-a}}{\sigma\sqrt{2\pi}}$$

where

$$a = \frac{(\mu - x)^2}{2\sigma^2}$$

Thus the normal distribution is defined in terms of two parameters, the mean (μ) and the standard deviation (σ) (Rees, 1987).

4. The Binormal Distribution

The binormal distribution is here defined as the sum of two normal distributions and is therefore a continuous probability distribution. The probability density function of the binormal distribution in variate X is given by:

$$P(X) = \rho.P(X_D) + (1-\rho).P(X_{\bar{D}})$$

It is a necessary feature of any probability density function that the area under the curve described by the function be equal to one. In order to preserve this condition the normal pdf of each subpopulation is multiplied by the prevalence of that subpopulation. Since the prevalence of the low attaining subpopulation is given by ρ, and the sum of the prevalences must equal one, it follows that the prevalence of the typical attainers must be equal to (1-ρ).

Thus the binormal distribution is thus defined in terms of five parameters. These are the mean and standard deviation of each subpopulation and the prevalence of the low attaining subpopulation $(\bar{x}_D, \bar{x}_{\bar{D}}, s_D, s_{\bar{D}}, \rho)$.

There is an intimate relationship between these parameters and the mean and standard deviation of the population overall such that:

$$\mu = \rho.\bar{x}_D + (1-\rho)\bar{x}_{\bar{D}} \qquad 1$$

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + (\bar{x}_D - \mu)(\mu - \bar{x}_{\bar{D}}) \qquad 2$$

The remainder of this chapter is concerned with providing an algebraic proof of these relationships.

5. Algebraic Proof of Equation 1 (Population Mean)

Total number of individuals in the population:

$$N = n_D + n_{\bar{D}} \tag{1.01}$$

Prevalence of the low attaining subpopulation:

$$\rho = \frac{n_D}{n_D + n_{\bar{D}}} \tag{1.02}$$

$$\frac{\rho}{n_D} = \frac{1}{n_D + n_{\bar{D}}} \tag{1.03}$$

Prevalence of the typically attaining subpopulation:

$$(1-\rho) = \frac{n_{\bar{D}}}{n_D + n_{\bar{D}}} \tag{1.04}$$

$$\frac{(1-\rho)}{n_{\bar{D}}} = \frac{1}{n_D + n_{\bar{D}}} \tag{1.05}$$

Population mean:

$$\mu = \frac{\sum X}{N} \tag{1.06}$$

Where:

$$\sum X = \sum X_D + \sum X_{\bar{D}} \tag{1.07}$$

Mean of low attaining subpopulation:

$$\bar{x}_D = \frac{\sum X_D}{n_D} \tag{1.08}$$

Mean of typically attaining subpopulation:

$$\bar{x}_{\bar{D}} = \frac{\sum X_{\bar{D}}}{n_{\bar{D}}} \tag{1.09}$$

Substituting 1.01 and 1.07 into 1.06:

$$\mu = \frac{\sum X_D + \sum X_{\bar{D}}}{n_D + n_{\bar{D}}} \quad 1.10$$

$$\mu = \frac{\sum X_D}{n_D + n_{\bar{D}}} + \frac{\sum X_{\bar{D}}}{n_D + n_{\bar{D}}} \quad 1.11$$

Substituting 1.03 and 1.05 into 1.11:

$$\mu = \frac{\rho.\sum X_D}{n_D} + \frac{(1-\rho).\sum X_{\bar{D}}}{n_{\bar{D}}} \quad 1.12$$

Substituting 1.08 and 1.09 into 1.12 gives an expression for the mean of the binormal distribution (equation 1):

$$\mu = \rho.\bar{x}_D + (1-\rho)\bar{x}_{\bar{D}}$$

6. Algebraic Proof of Equation 2 (Population Variance)

Population variance:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \qquad 2.01$$

Where:

$$\sum(X-\mu)^2 = \sum(X_D-\mu)^2 + \sum(X_{\overline{D}}-\mu)^2 \qquad 2.02$$

Variance of low attaining subpopulation:

$$s_D^2 = \frac{\sum(X_D-\overline{x}_D)^2}{n_D} \qquad 2.03$$

Variance of typically attaining subpopulation:

$$s_{\overline{D}}^2 = \frac{\sum(X_{\overline{D}}-\overline{x}_{\overline{D}})^2}{n_{\overline{D}}} \qquad 2.04$$

Substituting 1.01 and 2.02 into 2.01:

$$\sigma^2 = \frac{\sum(X_D-\mu)^2 + \sum(X_{\overline{D}}-\mu)^2}{n_D + n_{\overline{D}}} \qquad 2.05$$

$$\sigma^2 = \frac{\sum(X_D-\mu)^2}{n_D + n_{\overline{D}}} + \frac{\sum(X_{\overline{D}}-\mu)^2}{n_D + n_{\overline{D}}} \qquad 2.06$$

Substituting 1.03 and 1.05 into 2.06:

$$\sigma^2 = \frac{\rho.\sum(X_D-\mu)^2}{n_D} + \frac{(1-\rho).\sum(X_{\overline{D}}-\mu)^2}{n_{\overline{D}}} \qquad 2.07$$

Consider $(X_D-\mu)$:

$$(X_D-\mu) = (X_D-\overline{x}_D) + (\overline{x}_D-\mu) \qquad 2.08$$

$$(X_D - \mu)^2 = (X_D - \bar{x}_D)^2 + 2(X_D - \bar{x}_D)(\bar{x}_D - \mu) + (\bar{x}_D - \mu)^2 \quad 2.09$$

$$\sum(X_D - \mu)^2 = \sum(X_D - \bar{x}_D)^2 + \sum 2(X_D - \bar{x}_D)(\bar{x}_D - \mu) + \sum(\bar{x}_D - \mu)^2 \quad 2.10$$

Since $\sum(X_D - \bar{x}_D) = 0$:

$$\sum(X_D - \mu)^2 = \sum(X_D - \bar{x}_D)^2 + \sum(\bar{x}_D - \mu)^2 \quad 2.11$$

$$\frac{\sum(X_D - \mu)^2}{n_D} = \frac{\sum(X_D - \bar{x}_D)^2}{n_D} + \frac{\sum(\bar{x}_D - \mu)^2}{n_D} \quad 2.12$$

Substituting $\sum(\bar{x}_D - \mu)^2 = n_D(\bar{x}_D - \mu)^2$:

$$\frac{\sum(X_D - \mu)^2}{n_D} = \frac{\sum(X_D - \bar{x}_D)^2}{n_D} + (\bar{x}_D - \mu)^2 \quad 2.13$$

Substituting 2.03 into 2.13:

$$\frac{\sum(X_D - \mu)^2}{n_D} = s_D^2 + (\bar{x}_D - \mu)^2 \quad 2.14$$

And similarly:

$$\frac{\sum(X_{\bar{D}} - \mu)^2}{n_{\bar{D}}} = s_{\bar{D}}^2 + (\bar{x}_{\bar{D}} - \mu)^2 \quad 2.15$$

Substituting 2.14 and 2.15 into 2.07:

$$\sigma^2 = \rho.\left(s_D^2 + (\bar{x}_D - \mu)^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + (\bar{x}_{\bar{D}} - \mu)^2\right) \quad 2.16$$

Substituting Equation 1 into 2.16:

$$\sigma^2 = \rho.\left(s_D^2 + (\bar{x}_D - \rho.\bar{x}_D - (1-\rho)\bar{x}_{\bar{D}})^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + (\bar{x}_{\bar{D}} - \mu)^2\right) \quad 2.17$$

$$\sigma^2 = \rho.\left(s_D^2 + ((1-\rho).\bar{x}_D - (1-\rho)\bar{x}_{\bar{D}})^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + (\bar{x}_{\bar{D}} - \mu)^2\right) \quad 2.18$$

$$\sigma^2 = \rho.\left(s_D^2 + ((1-\rho)(\bar{x}_D - \bar{x}_{\bar{D}}))^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + (\bar{x}_{\bar{D}} - \mu)^2\right) \quad 2.19$$

Substituting Equation 1 into 2.19:

$$\sigma^2 = \rho.\left(s_D^2 + \left((1-\rho)(\bar{x}_D - \bar{x}_{\bar{D}})\right)^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + \left(\bar{x}_{\bar{D}} - \rho.\bar{x}_D - (1-\rho)\bar{x}_{\bar{D}}\right)^2\right) \quad 2.20$$

$$\sigma^2 = \rho.\left(s_D^2 + \left((1-\rho)(\bar{x}_D - \bar{x}_{\bar{D}})\right)^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + \left(\rho.\bar{x}_{\bar{D}} - \rho.\bar{x}_D\right)^2\right) \quad 2.21$$

$$\sigma^2 = \rho.\left(s_D^2 + \left((1-\rho)(\bar{x}_D - \bar{x}_{\bar{D}})\right)^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + \left(\rho(\bar{x}_{\bar{D}} - \bar{x}_D)\right)^2\right) \quad 2.22$$

$$\sigma^2 = \rho.\left(s_D^2 + (1-\rho)^2(\bar{x}_D - \bar{x}_{\bar{D}})^2\right) + (1-\rho).\left(s_{\bar{D}}^2 + \rho^2(\bar{x}_{\bar{D}} - \bar{x}_D)^2\right) \quad 2.23$$

Since $(\bar{x}_D - \bar{x}_{\bar{D}})^2 = (\bar{x}_{\bar{D}} - \bar{x}_D)^2$:

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(1-\rho)(\bar{x}_D - \bar{x}_{\bar{D}})^2 \quad 2.24$$

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(1-\rho)(\bar{x}_D - \bar{x}_{\bar{D}})(\bar{x}_D - \bar{x}_{\bar{D}}) \quad 2.25$$

Rearranging Equation 1:

$$\bar{x}_{\bar{D}} = \frac{\mu - \rho.\bar{x}_D}{1-\rho} \quad 2.26$$

Substituting 2.26 into 2.25:

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(1-\rho)\left(\bar{x}_D - \frac{\mu - \rho.\bar{x}_D}{1-\rho}\right)(\bar{x}_D - \bar{x}_{\bar{D}}) \quad 2.27$$

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(1-\rho)\left(\frac{\bar{x}_D - \rho.\bar{x}_D}{1-\rho} - \frac{\mu - \rho.\bar{x}_D}{1-\rho}\right)(\bar{x}_D - \bar{x}_{\bar{D}}) \quad 2.28$$

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(1-\rho)\left(\frac{\bar{x}_D - \mu}{1-\rho}\right)(\bar{x}_D - \bar{x}_{\bar{D}}) \quad 2.29$$

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(\bar{x}_D - \mu)(\bar{x}_D - \bar{x}_{\bar{D}}) \quad 2.30$$

Rearranging Equation 1:

$$\bar{x}_D = \frac{\mu - (1-\rho)\bar{x}_{\bar{D}}}{\rho} \quad 2.31$$

Substituting 2.31 into 2.30:

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(\bar{x}_D - \mu)\left(\frac{\mu - (1-\rho)\bar{x}_{\bar{D}}}{\rho} - \bar{x}_{\bar{D}}\right) \quad 2.32$$

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(\bar{x}_D - \mu)\left(\frac{\mu - (1-\rho)\bar{x}_{\bar{D}}}{\rho} - \frac{\rho.\bar{x}_{\bar{D}}}{\rho}\right) \quad 2.33$$

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + \rho.(\bar{x}_D - \mu)\left(\frac{\mu - \bar{x}_{\bar{D}}}{\rho}\right) \quad 2.34$$

Simplifying equation 2.34 gives an expression for the variance of the binormal distribution (equation 2):

$$\sigma^2 = \rho.s_D^2 + (1-\rho)s_{\bar{D}}^2 + (\bar{x}_D - \mu)(\mu - \bar{x}_{\bar{D}})$$

7. Summary

In this chapter the binormal distribution was defined as the continuous probability distribution resulting from the rescaled sum of two normal distributions. It was shown that the distribution is described by five parameters. The model was then developed algebraically to determine the relationships between these parameters. These relationships are summarised in equations that describe the population mean and population variance of the binormal distribution (equation 1 and equation 2). These equations are utilised in chapter 3 for the purpose of deriving estimates of the binormal parameters.

Chapter 3: A Method for Deriving Binormal Parameters

1. Introduction

In the previous chapter the binormal distribution was defined as a continuous probability distribution that can be described in terms of five parameters. In the present chapter a methodology is described for deriving estimates for these parameters from an empirical dataset. The reasons for deriving such parameters are threefold.

1. To provide evidence to support the hypothesis that the population consists of two qualitatively distinct subpopulations.
2. To describe in quantitative terms the relationship between those subpopulations.
3. To provide a basis for performing inferential statistics on the differences between the two subpopulations.

The methodology to derive binormal parameter estimates has two stages. In the first stage a curve was generated that described the cumulative probability distribution observed in the data. A curve fitting approach was then used to fit the binormal cumulative distribution function (cdf) to the observed curve. This technique employed nonlinear regression to determine the combination of parameter magnitudes that gave the best fit of the model to the observed data.

2. The Binormal Cumulative Distribution Function

The cumulative distribution function (cdf) is a curve that describes the cumulative probability distribution for a particular probability distribution. The cdf of the normal distribution is given by the following equation where *erf* is the Gauss error function.

$$\Phi(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

Since the binormal distribution is defined as the sum of two normal distributions, corrected for the prevalence of each subpopulation, it follows that the cdf of the binormal distribution is given by:

$$\Phi(x) = \left(\frac{\rho}{2}\right)\left(1 + erf\left(\frac{x_D - \bar{x}_D}{s_D\sqrt{2}}\right)\right) + \left(\frac{(1-\rho)}{2}\right)\left(1 + erf\left(\frac{x_{\bar{D}} - \bar{x}_{\bar{D}}}{s_{\bar{D}}\sqrt{2}}\right)\right)$$

3. Deriving the Observed Cumulative Probability Distribution

The cumulative probability distribution curve for a variate X is observed when the cumulative probability that a value is less than or equal to X is plotted against X. The value for X is taken directly from the data. The cumulative probability for each value of X in a data set containing N observations was derived by the following procedure.

1. The observations were sorted from lowest to highest and assigned the rank 1 to N.
2. The fractional rank for each observation was determined by dividing each rank by N, thus giving a scale running from 1/N to 1. Since the cumulative probability should properly run from 0 to 1 this method provides an estimate with a slight positive offset. The offset will be negligible where N is large, however the following steps were used to provide a simple correction.
3. The observations were sorted from highest to lowest and assigned the reverse rank. Thus the observation with the forward rank “1” was assigned the reverse rank “N”.
4. The reverse fractional rank was calculated for each observation by dividing by N.
5. A correction for the reversion was made by subtracting the reverse fractional rank from 1. This gave a scale that ran from 0 to (N-1)/N.
6. The cumulative probability for each observation was then calculated by finding the mean of the forward fractional rank and the corrected reverse fractional rank. This gave a cumulative probability scale that ran from 0 to 1.

In summary the cumulative probability for each observation is given by the following expression where *fFR* is the forward fractional rank and *rFR* is the reverse fractional rank.

$$\frac{fFR + (1 - rFR)}{2}$$

Having determined x and y coordinates for the observed cumulative probability distribution curve the next stage was to determine how well those data fitted the normal and binormal distribution models.

4. Curve Fitting

All curve fitting procedures were carried out using the DataFit software package (Oakdale Engineering, 2008). The software requires data to be input in the form of x and y coordinates. It has the capacity to program in user defined models and contains a variety of internal functions to facilitate this. The square root function (sqr) and Gauss error function (erf) were utilised to program cumulative distribution models for both the normal and binormal distributions.

The cumulative normal distribution model was simply coded as follows.

Model Definition:
F1 = (x-m)/(s*sqr(2))
Y = (1+erf(F1))/2

The model definition uses the letters ‘m’ and ‘s’ to represent the mean and standard deviation parameters of the normal distribution respectively.

Coding of the binormal distribution model was a little more complicated, as is explained below. The actual code used in the DataFit software was as follows.

Model Definition:
F1 = (m-p*m1)/(1-p)
F2 = sqr((s*s-p*s1*s1-(m1-m)*(m-(F1)))/(1-p))
F3 = (x-m1)/(s1*sqr(2))
F4 = (x-(F1))/((F2)*sqr(2))
F5 = (1+erf(F3))/2
F6 = (1+erf(F4))/2
Y = p*(F5)+(1-p)*(F6)

The binormal distribution is defined in terms of five parameters. In the model definition here employed parameters describing the distribution of the low attainment subpopulation are used directly thus; 'p' represents the prevalence, 'm1' is used to represent the subpopulation mean, and 's1' signifies the standard deviation of the same subpopulation. The two remaining parameters required to complete the model are the mean and standard deviation of the typically attaining subpopulation. However rather than entering them into the model definition directly, substitutions derived from the rearrangement of the equations for the population mean and variance of the binormal distribution that were derived in chapter 2 were used instead.

Rearranging the equation for the mean of the binormal distribution (equation 1) allows the mean of the typically attaining subpopulation to be expressed in terms of the population mean and the mean of the low attaining subpopulation.

$$\bar{x}_{\bar{D}} = \frac{\mu - \rho.\bar{x}_D}{1-\rho} \qquad 3$$

Similarly rearranging the equation for the variance of the binormal distribution (equation 2) allows the standard deviation of the typically attaining subpopulation to be expressed in terms of other parameters, including the standard deviation of the whole population.

$$s_{\bar{D}} = \sqrt{\frac{\sigma^2 - \rho.s_D^2 - (\bar{x}_D - \mu)(\mu - \bar{x}_{\bar{D}})}{1-\rho}} \qquad 4$$

The overall effect of these two substitutions is that two latent parameters in the model are replaced by two that can be calculated directly from the data, namely the population mean and standard deviation. As with the cumulative normal distribution, these are represented by ‘m’ and ‘s’ in the model definition.

To determine how well the observed data fitted each of the two models a nonlinear regression approach was used. Nonlinear regression requires that initial parameter estimates are entered into the software. DataFit then calculates how well the observed cumulative probability curve fits the model curve using those parameters. The software then enters an iterative procedure in which the parameter estimates are altered slightly with the aim of improving the overall model fit. This process continues until either a predetermined number of iterations have been completed or there is no discernable improvement in the model fit.

A potential difficulty with the nonlinear regression methodology is that it is possible to obtain more than one valid solution, depending on the values used for the initial parameter estimates. This is more likely to occur if the model uses a large number of parameters. In the case of the binormal distribution this is a distinct possibility since it is defined by five parameters. The solution to this is to replace variable parameters in the model with objectively determined constant values. The purpose of substituting latent parameters in the binormal model with the population mean and standard deviation was to facilitate this. Within the DataFit software it is possible to specify whether each variable in the model should be treated as a variable or constant value. In the present study the population mean and population standard deviation were assigned as constants with a value calculated from the data. Thus for the purpose of obtaining a model fit the normal distribution is expressed in terms of two constants, whilst the binormal distribution is expressed in terms of two constants and three variables.

5. Nonlinear Regression Output Statistics

The DataFit software produces a variety of output statistics that enable the user to evaluate their model fits. The three principle statistics that are reported in this study are outlined below.

The coefficient of multiple determination (R^2) reports the proportion of variation in the data points that is explained by the regression model. If the value of R^2 is equal to one it means that the curve passes through every data point. An R^2 value of zero means that the regression model does not describe the data any better than a horizontal line passing through the average y-value of the data points.

The adjusted coefficient of multiple determination (Ra^2) adjusts the value of R^2 according to the number of explanatory terms in the model used. The binormal distribution model includes three more variables than the simpler normal distribution model. It is not surprising therefore that it tends to provide a better fit to the data and higher R^2 values. The Ra^2 statistic adjusts the value of R^2 downwards to compensate for these additional degrees of freedom. Direct comparison of the Ra^2 statistic allows the evaluation of whether the improvement in the fit of the model justifies the increase in its complexity.

The nonlinear regression technique produces estimates of the magnitude of each variable in the model, together with the standard error of that estimate. These statistics are used to perform a t-test with the null hypothesis that the value of the variable is equal to zero. The prob(t) statistic is useful because it highlights any variables in the model that are not contributing to the overall fit and may therefore be dropped.

6. Limitations of the Methodology

A well known limitation of nonlinear regression, and one that is inherent in any iterative procedure, is the possibility of arriving at a false solution. This possibility increases with the complexity of the regression model. The possibility of arriving at a false solution may be compensated for by running the procedure a

number of times with different initial parameter estimates. Whatever the value of the initial parameter estimates the procedure will ideally converge upon the same solution. If more than one solution is found these may be evaluated to determine which, if any, is more likely to be the correct solution. More than one solution to the model fit may also indicate that there is greater complexity in the data than is explained by the regression model, or that a different model may be more appropriate.

A more serious limitation to the methodology concerns the quality of the data used to generate the observed cumulative probability distribution. It is a prerequisite that the data used provides a true reflection of the distribution of the abilities in the sample, and it follows that these must also be interval level data. Strictly speaking the standardised scores produced by traditional paper-based assessments are at the ordinal level. What is more, the pattern of scores on such assessments has as much to do with the distribution of item difficulties as it does with that of pupil abilities. If such data were used any findings generated by this methodology may simply be an artefact of the assessment used to gather the data. A partial solution to this is to employ a Rasch procedure to convert the data to interval level scores that truly reflect the distribution of abilities in the sample. However the use of Rasch measurement on its own is not sufficient. Paper-based (static) assessments are designed to target typically attaining individuals, with the result that those in the tails of the ability range are measured far less precisely. If the methodology is to provide good parameter estimates for a low attaining subpopulation it is necessary to get accurate ability measures in this range. This may be achieved using a dynamic procedure such as that offered by a computerised adaptive assessment. Using such an assessment system the precision with which an individual is measured is for the most part independent of where they fall on the ability range.

In addition to the quality of the data used it is also necessary to consider the quantity of those data. This affects in particular the fineness of the probability scale that forms the y-axis of the observed cumulative probability distribution. To illustrate this point consider the situation where there are only ten data points. These points are equally spaced in the vertical dimension at 0.1 intervals. The

figure of 0.1 represents the maximum possible precision of measurement in this dimension. Put another way the cumulative probability cannot be reported to any greater accuracy than one significant digit. Increasing the number of data points by a factor of ten to 100 would allow a theoretical reporting limit of two significant digits. However in practice we might require considerably more data-points before we were confident in reporting that level of accuracy. If it were taken as a 'rule-of-thumb' that the number of significant digits that we were confident in reporting were one less than the order of magnitude of the data points, and that a minimum acceptable level of precision were two significant digits, then the methodology described requires a minimum sample size of 1000.

7. Summary

In this chapter a methodology is described for deriving binormal parameter estimates from an empirical dataset. The procedure is carried out in two phases. In the first phase the coordinates of data points that describe the cumulative probability curve observed in the dataset are generated. In the second phase nonlinear regression is used to determine the combination of parameter estimates that provide the closest fit of the observed data to the theoretical model. It is argued that both the quality and quantity of data are important considerations in the application of the methodology. It is recommended that the data used should be collected using an adaptive assessment procedure, with a minimum sample size of one thousand individuals.

Chapter 4: The Data Used to Evaluate the Binormal Distribution Model

1. Introduction

Having defined the binormal distribution model and developed a methodology for deriving parameter estimates for the same, the next stage was to apply the model to a real dataset. Given the limitations of the methodology described in the previous chapter careful consideration was given to the data chosen to evaluate the model. A substantial database of assessment results was available that had been gathered using an adaptive computer based system, thus meeting the necessary quality and quantity criteria. In this chapter a description is given of the assessment used, and of the data collected. Finally, the procedure and rationale for preparing the raw data for input into the regression model is also described.

2. The Interactive Computerised Assessment System

The Interactive Computerised Assessment System (InCAS) is a computer adaptive assessment designed by the Centre for Evaluation and Monitoring at Durham University, UK. The InCAS software provides a collection of assessment modules that are designed for use by children aged between five and eleven years. Each of the assessment items that InCAS uses are thoroughly trialled in advance to determine the age at which a typically achieving child would have an equal chance of answering it correctly or incorrectly. This age represents the difficulty of that item. The software then takes the child's chronological age as an initial estimate of their ability and presents the first item accordingly. Typically this would mean presenting the child with an item difficulty two years lower than their chronological age. InCAS then uses adaptive algorithms to select the difficulty of the items that are subsequently presented. In this way the software quickly targets items to the ability of the child, greatly increasing the efficiency and reliability of the assessment process. In the same way that the item difficulties are defined in terms of an age-equivalency, so are the ability measures that are output by the software. A fuller description of

InCAS and the rationale behind the assessment methodology has been given by Merrell and Tymms (Merrell and Tymms, 2007).

For the purposes of the present study data from four InCAS assessment modules were used. These were picture vocabulary, reading, mathematics and arithmetic. The format of each of the modules is described below. Example screenshots from each assessment module are given in the appendix.

Picture Vocabulary: This is a relatively simple task in which the child hears a word and sees five pictures. They use the mouse pointer to click on the picture that best illustrates that word.

Reading: The InCAS reading module itself consists of three separate tasks. These are word recognition, word decoding and reading comprehension. In the word recognition task the child hears a word, and then a sentence putting that word into context. They use the mouse pointer to click on the correct spelling of that word from a choice of five. Clearly children may use both word recognition and word decoding strategies to correctly answer each question. In order to disentangle these two quite different skills they are then presented with a dedicated word decoding task. This takes a similar format to the word recognition task, but in this case the child is presented with an unfamiliar or nonsense word. Since the child will not have seen the word before they must use a decoding strategy to find the correct solution. Taken together the word recognition and decoding tasks provide a measure of basic reading skills. This is used to select a passage of text that is of suitable difficulty for the reading comprehension task. In this final reading task the child is asked at various points to select the correct word from a choice of three that best fits within the overall meaning of the sentence.

Mathematics: The mathematics module covers the broad maths curriculum including number, measurement, shape & space, and data handling. The child hears a question and is given additional visual information in the form of pictures, charts etc. The child then selects the correct answer from a choice of up to five.

Arithmetic: The arithmetic module consists of four tasks, one for each arithmetic operation. Progression through the tasks from addition to subtraction, multiplication, and finally division depends on the performance on the previous arithmetic tasks. Each assessment item is presented visually in symbolic notation with a choice of four possible answers to click the mouse pointer on.

Reliability figures were calculated for each module and were as follows; picture vocabulary (0.89), reading (0.97), mathematics (0.97), and arithmetic (0.96). The validity of the picture vocabulary and reading modules has been discussed by Merrell and Tymms (Merrell and Tymms, 2007). For these modules predictive validity was determined by comparison with PIPS paper based standardised assessments. The correlations between the PIPS and InCAS assessments were found to be 0.82 for picture vocabulary and 0.75 for reading. Both figures were statistically significant at the 0.01 level. The mathematics module was developed using items from a variety of sources, including well validated items released by the Third International Mathematics and Science Study (IEA, 1995a, IEA, 1995b). The item difficulties reported by the IEA were compared with those that were generated when the same items were presented using the InCAS assessment engine. The correlations were found to be 0.79 (30 items; population 1) and 0.77 (15 items; population 2). Both figures were statistically significant at the 0.01 level. No figures were available for the validity of the InCAS arithmetic module.

3. The Assessment Sample

The reading and maths modules of the InCAS assessment are a statutory requirement for all state schools in Northern Ireland during the last four years of primary school (that is P4, P5, P6 and P7). The assessments are taken by all of the children in these year groups except in very special circumstances. In addition to the statutory requirements many schools undertake the other assessment modules such as picture vocabulary and arithmetic on a voluntary basis. The statutory nature of the maths and reading modules means that data are available on a substantial majority of the population in those four year groups,

and data were gathered from in excess of 80,000 children. The voluntary modules were completed by about a quarter the eligible population.

The data presented here were collected in Northern Ireland during 2009. This particular dataset was chosen for two reasons. Although the InCAS assessment has been a statutory requirement in Northern Ireland since 2007, it was previously restricted to one or two cohorts. This was the first occasion that data were available across the four upper primary cohorts. In addition there had been significant development of InCAS since 2008 to lower the floor and raise the ceiling of both the reading and mathematics assessments, thus extending the ability range over which reliable measurement could be made. All of the assessments were completed between September and December, with peak activity around October. To the author's knowledge the data collected are unique with respect to quantity for a computerised adaptive assessment.

4. The Assessment Process

The assessments were carried out by teachers and/or teaching assistants within the participating schools. The children were typically assessed in small to whole class groups according to the group dynamics, available staff and availability of computers. Specific details of the management of the process were left to each school.

From the child's perspective the assessment process involves sitting at the computer terminal wearing a set of headphones in order to hear the audio component of the assessment. Initiation of each assessment is controlled through the use of unique passwords that are used only once. This system facilitates management of the process by school staff and helps to ensure that each child is assessed once only on each module. Once a password is entered the child is given some audio instructions on how to complete the assessment, followed by some practice questions. The assessment itself proceeds without providing feedback to the child on whether each item was answered correctly or not. A status bar provides an indication to the child of how much time remains to complete the assessment. InCAS has a time limit for each question and for the assessment overall, however these are generous and the assessment reaches a natural conclusion for the vast majority of children. If a child fails to respond to a particular item within the time limit it is recorded by the software as *'timed out'* but is treated as incorrect for the purposes of producing a final score. The time it takes to complete an assessment varies according to which module it is and the individual child, but most are completed within twenty minutes. It is recommended that children sit no more than one module in a single session, although modules such as arithmetic and reading contain a series of subtasks.

As each assessment proceeds InCAS records the item presented together with the child's response (correct, incorrect or timed out). At the end of the assessment InCAS calculates an overall score and separate scores for any subscale within the module. These data together with item level information are uploaded via a secure internet connection to the InCAS web server located at Durham University. It is these uploaded data that were used in the present study.

5. Initial Data Processing

The summary statistics generated by InCAS were disregarded in this study. Instead the item level responses were taken and used to generate the same statistics from scratch. The advantage of this approach was that the pupil scores were referenced directly within the concurrent sample, rather than on an item standardisation based on data collected in the previous year.

Interval level scores were generated from the item level data using the Winsteps Rasch-Model Computer Program (Linacre, 2007). The interval scale produced by Winsteps is in the form of logits. These were mapped onto an age scale using a straightforward linear regression of each child's logit score against their age at test. Summary statistics for these age equivalencies, broken down by year group are shown in tables 1.1 to 1.4.

A glance at tables 1.1 to 1.4 reveals that the mean age equivalent scores are broadly, but not exactly in line with the mean age at test. The reading module generated the most variable scores, and this was probably due to the greater complexity of the assessment, composed as it was of three separate tasks. The arithmetic module scores were more variable than both the mathematics or picture vocabulary scores, and this assessment also had a relatively complex structure of four subtasks. The variation of the reading module scores was fairly uniform across the four year groups. However in the remaining modules there was a noticeable increase in the variation in P7.

The pattern of assessment scores across cohorts was further investigated by dividing each year group into six subgroups according to age. Thus each subgroup contained children that were born within about two months of one another. The mean assessment score was then plotted against the mean age at test for each subgroup. This reveals the cross sectional pattern in assessment scores according to age and year group (grade), and are here referred to age-grade curves (figures 1.1 to 1.4).

Year Group	Number of Children	Mean Age at Test (years)	Age Equivalent Score (years)	
			Mean	Standard Deviation
P4	4028	7.80	7.67	1.87
P5	5151	8.79	8.83	1.81
P6	5399	9.80	9.86	1.88
P7	5384	10.79	10.80	2.03

Table 1.1: Summary statistics for the InCAS picture vocabulary module.

Year Group	Number of Children	Mean Age at Test (years)	Age Equivalent Score (years)	
			Mean	Standard Deviation
P4	18343	7.78	7.63	1.60
P5	19722	8.77	8.83	1.64
P6	20656	9.78	9.89	1.66
P7	21673	10.77	10.74	1.64

Table 1.2: Summary statistics for the InCAS reading module.

Year Group	Number of Children	Mean Age at Test (years)	Age Equivalent Score (years)	
			Mean	Standard Deviation
P4	18249	7.78	7.75	1.07
P5	19321	8.77	8.78	1.11
P6	20281	9.78	9.68	1.14
P7	21367	10.77	10.90	1.48

Table 1.3: Summary statistics for the InCAS mathematics module.

Year Group	Number of Children	Mean Age at Test (years)	Age Equivalent Score (years)	
			Mean	Standard Deviation
P4	4451	7.79	7.69	1.26
P5	5655	8.78	8.81	1.25
P6	5957	9.79	9.80	1.26
P7	6090	10.78	10.82	1.47

Table 1.4: Summary statistics for the InCAS arithmetic module.

Each age-grade curve shows a characteristic step up between one year group and the next. This step up shows the effect of one years schooling. In the case of the picture vocabulary and reading modules the step up is a little less each year, indicating in absolute terms a decrease in the effect of a years schooling as the children get older. The same pattern holds with mathematics and arithmetic for the first two steps. However the final step between P6 and P7 shows a marked increase in magnitude that goes against the general trend. There is also a trend within each year group favouring higher scores for older children. The slope within each year group decreases with older cohorts, indicating that the importance of age on assessment scores decreases as the children get older. This pattern is maintained across all four assessment modules.

It was clear from this preliminary inspection of the assessment results that the distribution of scores is influenced both by the number of years of schooling and the age of the child within each cohort. As a result it was decided that any treatment of the data be conducted separately for each year group. In addition it was decided that all further analyses be conducted with assessment scores that were corrected for age. This was achieved by simply subtracting each child's age at test from their age-equivalent score on each assessment module. The corrected score is here referred to as the age/ability difference.

Figure 1.1: Age-Grade Curve for the InCAS Picture Vocabulary Module

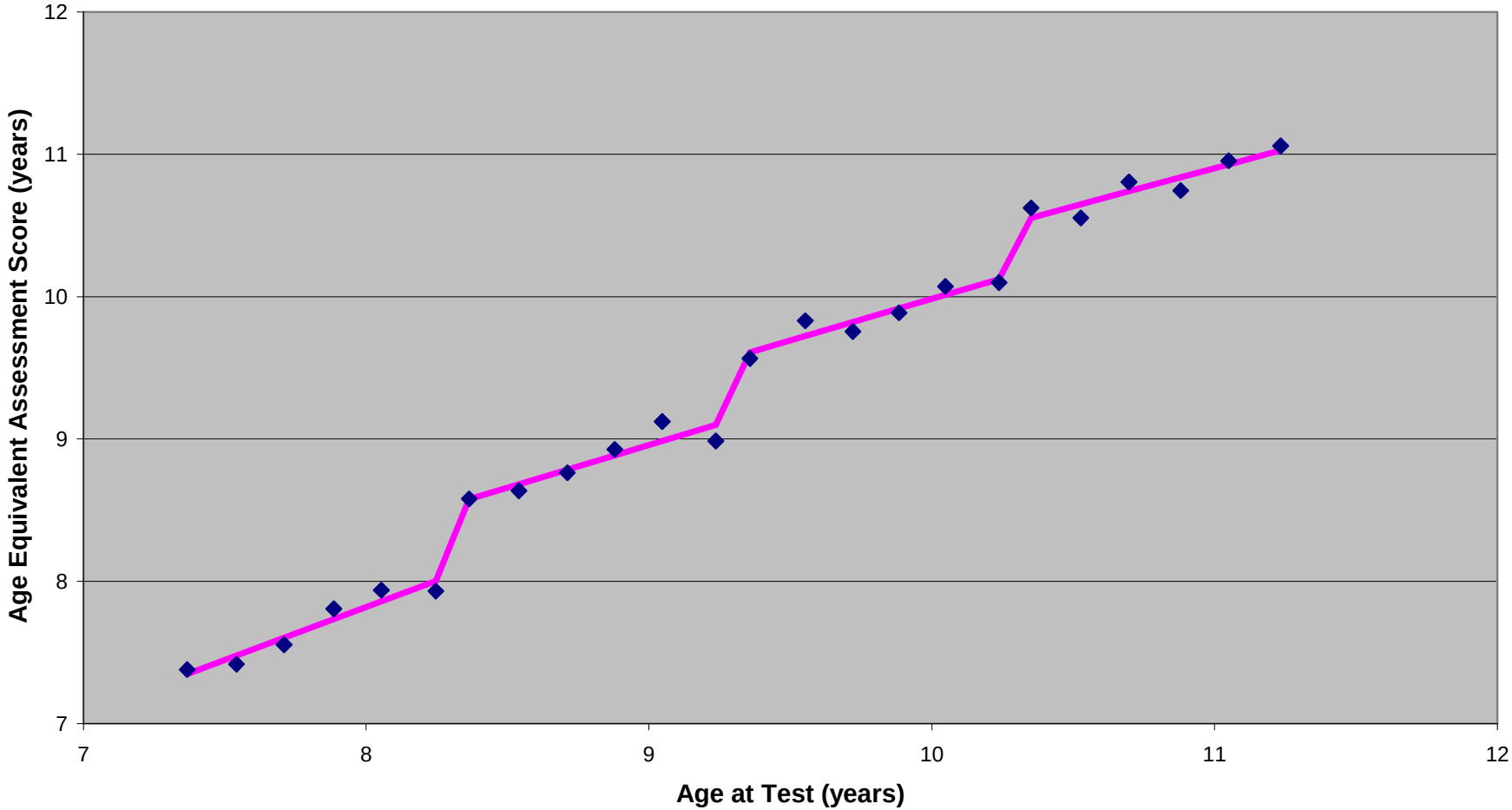


Figure 1.2: Age-Grade Curve for the InCAS Reading Module

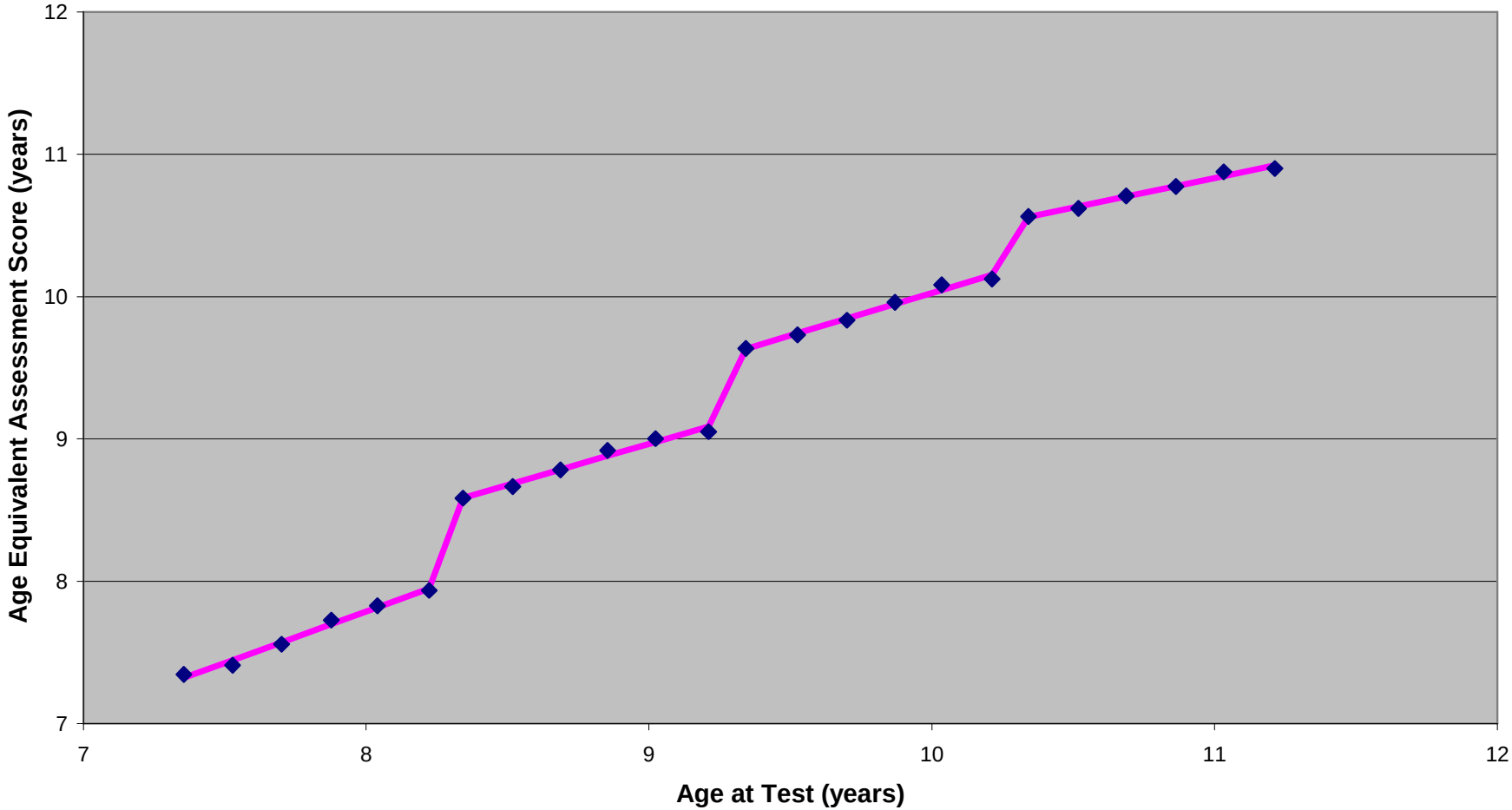


Figure 1.3: Age-Grade Curve for the InCAS Mathematics Module

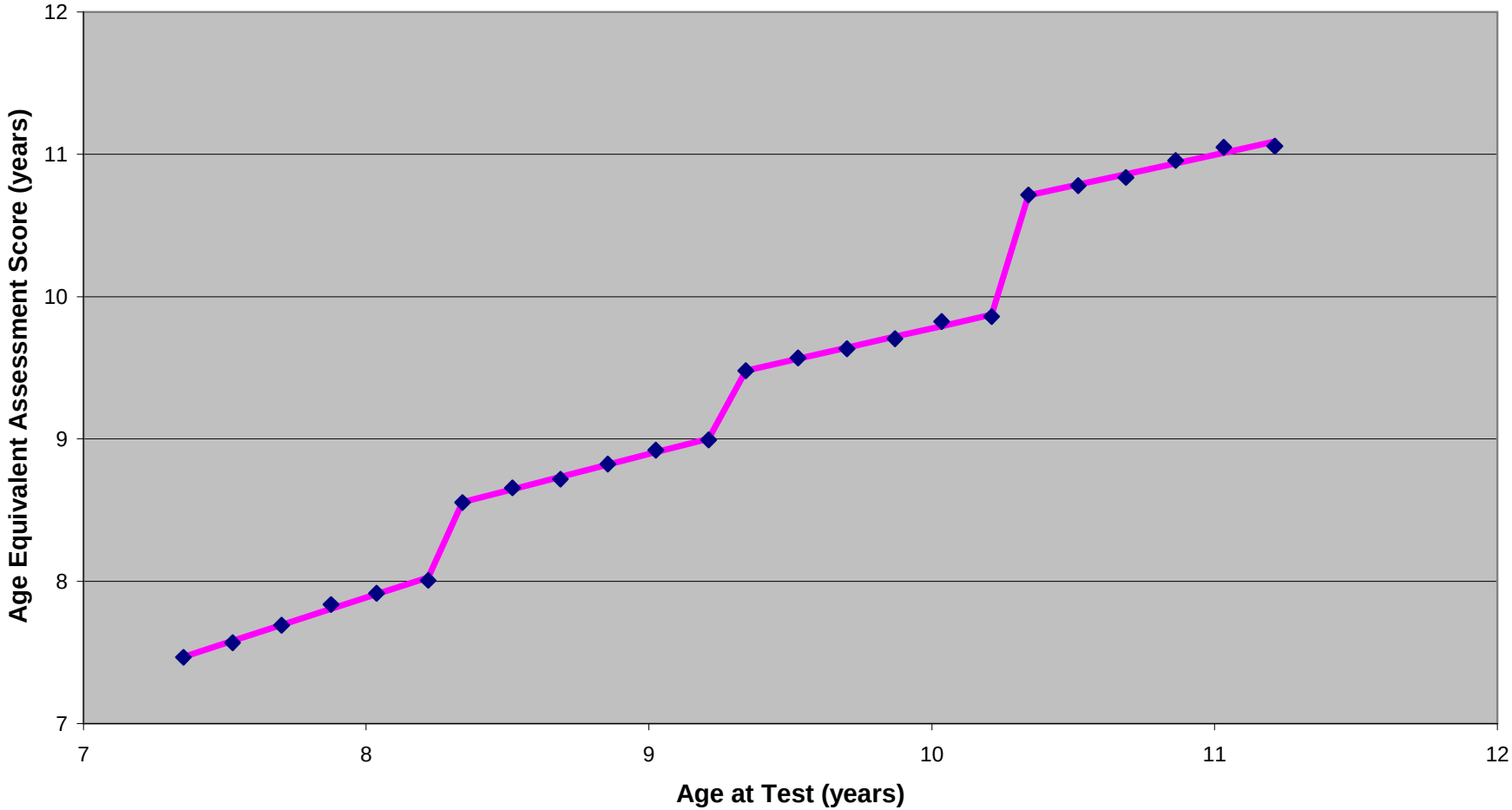
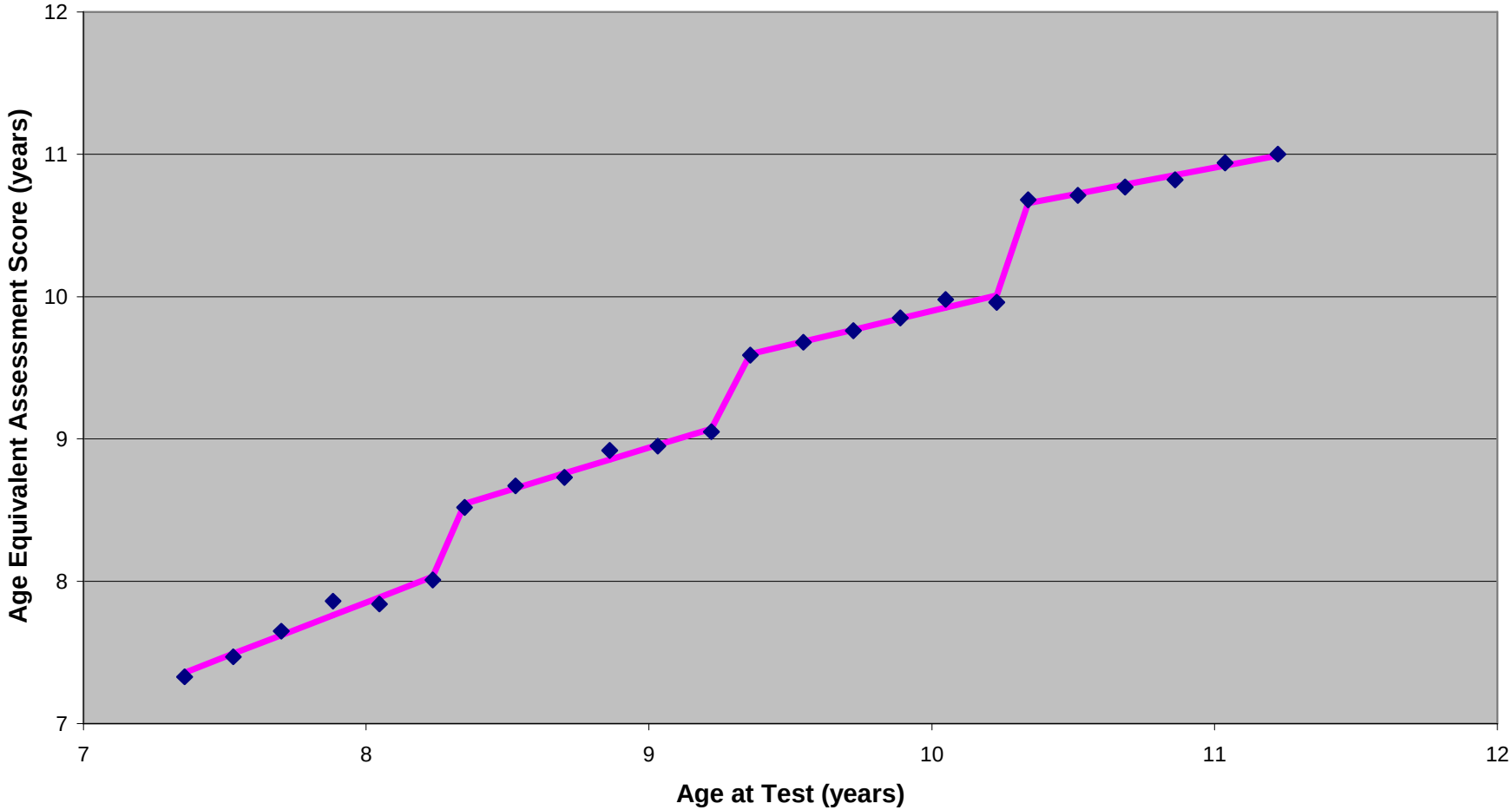


Figure 1.4: Age-Grade Curve for the InCAS Arithmetic Module



6. The Effect of Gender on the Assessment Scores

An issue of concern to the developers of educational assessments is that of gender bias. Assessments must at least be perceived by the teachers that use them not to favour either boys or girls. Of course differences between boys and girls in the pattern of scores may be due to genuine differences that arise from a complex interaction of a variety of causes, rather than as an artefact of the assessment materials and procedure. One relatively consistent feature of many assessments is that boys' scores tend to have a greater variation than those of girls. In the present context any gender bias that may be present in the InCAS assessment modules is not of any immediate concern because the study is primarily concerned with the shape of the distribution of scores rather than the magnitude of those scores. In the previous section it was argued that the year of schooling and the relative age of a child within a cohort are two factors that affect the shape of the score distribution, and that these can readily be taken into account in the analysis. In this section attention is focussed on the effect of gender.

Tables 2.1.1 to 2.4.2 present descriptive statistics for the age/ability difference in InCAS assessment module scores according to year group and gender. The results of significance tests for the equality of means and variances of these scores between boys and girls are given in Table 3 and Table 4 respectively.

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	2089	7.80	-0.173	1.957
P5	2649	8.80	0.035	1.850
P6	2730	9.80	0.154	1.917
P7	2736	10.78	0.128	2.096

Table 2.1.1: Descriptive statistics for boys' picture vocabulary results

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	1939	7.80	-0.088	1.745
P5	2502	8.79	0.038	1.757
P6	2669	9.80	-0.032	1.830
P7	2648	10.80	-0.130	1.977

Table 2.1.2: Descriptive statistics for girls' picture vocabulary results

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	9331	7.79	-0.386	1.656
P5	10166	8.77	-0.156	1.718
P6	10648	9.78	-0.072	1.724
P7	11149	10.77	-0.240	1.701

Table 2.2.1: Descriptive statistics for boys' reading results

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	9012	7.78	0.089	1.481
P5	9556	8.77	0.293	1.507
P6	10008	9.78	0.312	1.551
P7	10524	10.78	0.180	1.554

Table 2.2.2: Descriptive statistics for girls' reading results

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	9266	7.79	-0.034	1.108
P5	10011	8.77	-0.002	1.162
P6	10442	9.78	-0.101	1.203
P7	10964	10.77	0.107	1.542

Table 2.3.1: Descriptive statistics for boys' mathematics results

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	8983	7.78	-0.040	1.000
P5	9310	8.77	0.014	1.049
P6	9839	9.78	-0.103	1.093
P7	10403	10.78	0.142	1.419

Table 2.3.2: Descriptive statistics for girls' mathematics results

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	2254	7.79	-0.016	1.292
P5	2840	8.78	0.051	1.346
P6	3023	9.80	0.035	1.332
P7	3102	10.77	0.011	1.600

Table 2.4.1: Descriptive statistics for boys' arithmetic results

Year Group	Number of Pupils	Mean Age (years)	Age/Ability Difference (years)	
			Mean	Std Deviation
P4	2197	7.79	-0.187	1.173
P5	2815	8.79	0.000	1.131
P6	2934	9.79	-0.025	1.200
P7	2988	10.78	0.078	1.341

Table 2.4.2: Descriptive statistics for girls' arithmetic results

Year Group	Comparison of Means (significance)			
	Vocabulary	Reading	Mathematics	Arithmetic
P4	0.15	0.00	0.70	0.00
P5	0.96	0.00	0.31	0.13
P6	0.00	0.00	0.91	0.07
P7	0.00	0.00	0.09	0.08

Table 3: Comparison of mean age/ability difference scores by gender using an independent-samples t-test with equal variances not assumed.

Year Group	Comparison of Variances (significance)			
	Vocabulary	Reading	Mathematics	Arithmetic
P4	0.00	0.00	0.00	0.00
P5	0.00	0.00	0.00	0.00
P6	0.00	0.00	0.00	0.00
P7	0.20	0.00	0.00	0.00

Table 4: Comparison of the spread of age/ability difference scores by gender using Levene's test for homogeneity of variances.

The figures presented demonstrate that on average girls perform significantly better than boys on the reading assessment across all year groups, however in mathematics no such difference was observed. The boys' average arithmetic scores were significantly higher than those of the girls in P4, however this advantage had been lost by P5. By the time the children reached P7 the girls were slightly ahead of the boys in arithmetic, although the difference was not statistically significant. In terms of average assessment score the results for the picture vocabulary module were the most curious. In the earlier year groups, P4 and P5, there was no significant difference between boys and girls. However by P6 the boys were achieving significantly better results, and by P7 the gap had widened even further.

In terms of equality of variance it was found that boys' scores were significantly more variable than those of girls in almost every circumstance. The one exception to this was found in the case of P7 picture vocabulary where no evidence for a difference was observed.

Given these results it is clear that in every combination of assessment module and year group there is evidence for a difference in the distribution of scores according to gender. In consequence it was decided that application of the binormal model was to be conducted separately for boys and girls.

7. Summary

In this chapter a description is given of the method and materials used in the collection of data that would be used to evaluate the binormal distribution model. That is followed by an account of how the raw data were processed and descriptive statistics of those data. It is argued that the distribution of scores may be influenced by three. These are the year of schooling, the relative age within the cohort, and the sex of the child. Each of these factors was taken into account in the application of the methodology for deriving binormal parameters. The relative age within the cohort was allowed for by using age corrected scores. Year of schooling and the sex of the child were allowed for by analysing each grouping separately. A disadvantage of this approach is that the smaller size of the datasets thus used will tend to compromise the power of the methodology to produce accurate parameter estimates as discussed in chapter 3.

Chapter 5: A Statistical Evaluation of the Regression Model Fits

1. Introduction

In this chapter precise details are first given on how the methodology for deriving binormal parameter estimates was applied to the InCAS assessment data described in the previous chapter. This is followed by an initial evaluation of the model fits so derived. This initial evaluation was concerned with whether the relatively complex binormal model provided a better fit for the observed data than the default normal distribution model, and centres on the output statistics generated by the DataFit software.

2. Initial Parameter Estimates

As stated previously, a difficulty with the nonlinear regression technique is that it is possible to arrive at more than one solution, and that the chance of this increases with the number of variables in the regression model. One way of decreasing the number of variables in the model is to replace them with constant values. In the present study these constants were provided by the mean and standard deviation that were calculated directly from the data. The values of the parameter constants used are reported in tables 2.1.1 to 2.4.2.

In the case of the normal distribution the replacement of two variable parameters with two constant parameters in the model effectively fixed the model solution. In the case of the binormal distribution the same replacement left three variable parameters, and therefore the possibility of more than one solution to the model fit. For this reason two conditions of initial parameter estimates were used to evaluate each binormal regression model. These initial conditions are set out in table 5.

Variable Parameter	Initial Estimate #1	Initial Estimate #2
Prevalence	0.5	0.01
Mean	-0.5	-4.0
Standard Deviation	σ	σ

Table 5: Initial variable parameter estimates used to evaluate the binormal regression model under two conditions. The parameters refer to the low attaining subpopulation.

In each condition the initial estimate of the standard deviation of the low attaining subpopulation was taken as the standard deviation of the population overall. Since the mean age/ability difference was close to zero under all conditions the initial parameter estimates under condition #1 gave a scenario that was approximately symmetrical about the population mean. The initial parameter estimates under condition #2 reflected the scenario where there was a very small subpopulation of low attaining children in the far left-hand tail of the distribution.

In most cases the nonlinear regression converged to the same solution under both conditions of initial parameter estimates. In some cases under condition #2 a solution was reached in which either the prevalence or standard deviation had a negative value. In these circumstances the theoretically impossible solution was rejected in favour of the solution derived using the initial parameter estimates under condition #1. In cases where two different but theoretically possible solutions were reached the model fit reported is the one with the higher value for the coefficient of multiple determination (R^2), that is the model fit that explained more of the variation in the observed data.

3. The Overall Goodness of Fit

We should only accept the binormal distribution if it provides a better fit than the normal distribution after taking into account the greater flexibility in the model afforded by three additional parameters. Since the normal model is nested within the binormal model the null hypothesis that the binormal model does not give a significantly better fit can be evaluated using an F test. The value of F is calculated using the following expression where the subscripts $_1$ and $_2$ refer to the normal and binormal distributions respectively, *RSS* refers to the residual sum of squares of the model fit, *p* refers to the number of parameters that describe the model, and *n* is the number of data points.

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\left(\frac{RSS_2}{n - p_2}\right)}$$

The DataFit software reports the residual sum of squares as part of its output, and it is therefore relatively straightforward to calculate F statistics. The value of F thus calculated has (p_2 - p_1, n - p_2) degrees of freedom. When the F test was performed a probability of 0.00 was returned under all conditions. However this was due at least in part to the very large number of data points. The high value of *n* in these calculations made it extremely difficult to reject the null hypothesis, and so on these criteria at least it was accepted that the binormal distribution always gave a significantly better fit.

Another approach to evaluating the overall goodness of fit is to compare the adjusted coefficient of multiple determination (Ra^2). This statistic is a version of the coefficient of multiple determination that is adjusted to account for the number of variables in the regression model. This means that the fit statistics may be compared directly with one another. An advantage of using this method over using an F test is that the difference in Ra^2 statistics provides a quantitative indication of the improvement in fit. A comparison of Ra^2 fit statistics for the normal and binormal distribution models are given in tables 6.1.1 to 6.4.2.

The difference in Ra^2 when the value for the binormal model is subtracted from the value for the normal model is positive in all circumstances. This indicates that the binormal model always provides a statistically significant improvement in fit after allowing for the additional degrees of freedom in the model. This confirms the results of the F test described earlier in this section.

A cursory analysis of the magnitude in the difference in Ra^2 statistics indicates that the improvement in fit is least apparent in the picture vocabulary assessment, followed by reading and mathematics. The arithmetic assessment tends to show the most marked improvement in fit. With respect to the reading and mathematics modules the improvement in model fit tends to increase with the age of the cohort. An exception to this general trend is found in the mathematics results for both boys and girls where there was a decrease in the improvement of fit in P7.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.99453	0.99974	0.00521
P5	0.99570	0.99980	0.00410
P6	0.99471	0.99987	0.00516
P7	0.99586	0.99990	0.00403

Table 6.1.1: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to boys' picture vocabulary scores.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.99399	0.99966	0.00568
P5	0.99068	0.99991	0.00923
P6	0.99478	0.99956	0.00478
P7	0.99771	0.99944	0.00174

Table 6.1.2: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to girls' picture vocabulary scores.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.99956	0.99994	0.00038
P5	0.99478	0.99990	0.00512
P6	0.98976	0.99980	0.01004
P7	0.98637	0.99926	0.01289

Table 6.2.1: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to boys' reading scores.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.99931	0.99986	0.00055
P5	0.99587	0.99985	0.00398
P6	0.99436	0.99979	0.00543
P7	0.98847	0.99953	0.01106

Table 6.2.2: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to girls' reading scores.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.99868	0.99981	0.00113
P5	0.98912	0.99991	0.01079
P6	0.99001	0.99991	0.00990
P7	0.99515	0.99966	0.00451

Table 6.3.1: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to boys' mathematics scores.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.99828	0.99995	0.00166
P5	0.98955	0.99993	0.01038
P6	0.98887	0.99995	0.01108
P7	0.99372	0.99979	0.00607

Table 6.3.2: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to girls' mathematics scores.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.98482	0.99988	0.01505
P5	0.98234	0.99982	0.01748
P6	0.98887	0.99984	0.01097
P7	0.99495	0.99986	0.00491

Table 6.4.1: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to boys' arithmetic scores.

Year Group	Adjusted Coefficient of Multiple Determination		Difference in Ra^2 (Binormal – Normal)
	Normal Model	Binormal Model	
P4	0.97118	0.99986	0.02868
P5	0.98111	0.99979	0.01868
P6	0.97876	0.99979	0.02103
P7	0.97463	0.99989	0.02527

Table 6.1.1: Comparison of the adjusted coefficient of multiple determination (Ra^2) for normal and binormal models applied to girls' arithmetic scores.

4. Binormal Parameter Fit Statistics

The output from the DataFit software includes estimates of the value of any variable parameters in the model, together with the standard error of those estimates. It then performs a t-test and calculates the probability that the parameter estimate is actually zero. If the null hypothesis is accepted it is an indication that the parameter in question does not contribute to the overall model fit, and therefore that a simpler model requiring fewer parameters is more appropriate. If such an analysis were to result in the rejection of the binormal model then by default the normal model is accepted as the more appropriate.

As a result of the way in which the binormal model was coded into the DataFit software, direct parameter estimates were only generated for the prevalence, mean and standard deviation of the low attaining subpopulation. Parameter estimates for the higher attaining subpopulation were then calculated using equations 3 and 4. These data are presented in tables 7.1.1.1 to 7.4.4.2.

These results indicate that in every circumstance with the exception of the picture vocabulary scores for girls in P7 (table 7.1.4.2) the three variable parameters make a statistically significant contribution to the binormal model fit. In most circumstances the contribution is highly significant, but it is only just significant in the case of mathematics scores for boys in P4 (table 7.3.1.1).

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.073	0.021	3.515	0.000
$\bar{x}_D$	-3.272	0.776	-4.219	0.000
s_D	2.496	0.287	8.684	0.000
$\bar{x}_{\bar{D}}$	0.072			
$s_{\bar{D}}$	1.679			

Table 7.1.1.1: Binormal parameter estimates for boys' picture vocabulary scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.269	0.046	5.823	0.000
$\bar{x}_D$	-1.453	0.273	-5.325	0.000
s_D	1.908	0.067	28.375	0.000
$\bar{x}_{\bar{D}}$	0.415			
$s_{\bar{D}}$	1.373			

Table 7.1.1.2: Binormal parameter estimates for girls' picture vocabulary scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.534	0.008	70.703	0.000
$\bar{x}_D$	-0.367	0.009	-39.348	0.000
s_D	2.157	0.005	430.423	0.000
$\bar{x}_{\bar{D}}$	0.497			
$s_{\bar{D}}$	1.269			

Table 7.1.2.1: Binormal parameter estimates for boys' picture vocabulary scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.296	0.003	90.170	0.000
$\bar{x}_D$	-0.704	0.009	-76.851	0.000
s_D	2.422	0.005	459.664	0.000
$\bar{x}_{\bar{D}}$	0.350			
$s_{\bar{D}}$	1.261			

Table 7.1.2.2: Binormal parameter estimates for girls' picture vocabulary scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.229	0.007	31.516	0.000
$\bar{x}_D$	-1.115	0.043	-25.825	0.000
s_D	2.426	0.005	514.424	0.000
$\bar{x}_{\bar{D}}$	0.531			
$s_{\bar{D}}$	1.548			

Table 7.1.3.1: Binormal parameter estimates for boys' picture vocabulary scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.030	0.002	14.871	0.000
$\bar{x}_D$	-5.006	0.204	-24.602	0.000
s_D	1.697	0.181	9.397	0.000
$\bar{x}_{\bar{D}}$	0.120			
$s_{\bar{D}}$	1.608			

Table 7.1.3.2: Binormal parameter estimates for girls' picture vocabulary scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.279	0.008	37.010	0.000
$\bar{x}_D$	-0.973	0.033	-29.250	0.000
s_D	2.545	0.005	550.151	0.000
$\bar{x}_{\bar{D}}$	0.554			
$s_{\bar{D}}$	1.668			

Table 7.1.4.1: Binormal parameter estimates for boys' picture vocabulary scores in P7

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.175	0.194	0.900	0.368
$\bar{x}_D$	-1.799	1.825	-0.986	0.324
s_D	2.164	0.426	5.085	0.000
$\bar{x}_{\bar{D}}$	0.223			
$s_{\bar{D}}$	1.740			

Table 7.1.4.2: Binormal parameter estimates for girls' picture vocabulary scores in P7

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.763	0.005	165.795	0.000
$\bar{x}_D$	-0.872	0.009	-101.775	0.000
s_D	1.472	0.003	518.723	0.000
$\bar{x}_{\bar{D}}$	1.179			
$s_{\bar{D}}$	1.180			

Table 7.2.1.1: Binormal parameter estimates for boys' reading scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.366	0.052	7.049	0.000
$\bar{x}_D$	-0.817	0.128	-6.358	0.000
s_D	1.373	0.028	49.204	0.000
$\bar{x}_{\bar{D}}$	0.613			
$s_{\bar{D}}$	1.274			

Table 7.2.1.2: Binormal parameter estimates for girls' reading scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.166	0.002	99.677	0.000
$\bar{x}_D$	-2.531	0.013	-188.817	0.000
s_D	1.254	0.006	195.445	0.000
$\bar{x}_{\bar{D}}$	0.318			
$s_{\bar{D}}$	1.370			

Table 7.2.2.1: Binormal parameter estimates for boys' reading scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.146	0.007	20.419	0.000
$\bar{x}_D$	-1.631	0.067	-24.190	0.000
s_D	1.414	0.024	59.119	0.000
$\bar{x}_{\bar{D}}$	0.622			
$s_{\bar{D}}$	1.255			

Table 7.2.2.2: Binormal parameter estimates for girls' reading scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.163	0.002	88.315	0.000
$\bar{x}_D$	-2.618	0.018	-144.588	0.000
s_D	1.307	0.010	135.782	0.000
$\bar{x}_{\bar{D}}$	0.424			
$s_{\bar{D}}$	1.306			

Table 7.2.3.1: Binormal parameter estimates for boys' reading scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.081	0.001	70.485	0.000
$\bar{x}_D$	-2.571	0.021	-122.889	0.000
s_D	1.116	0.014	79.102	0.000
$\bar{x}_{\bar{D}}$	0.568			
$s_{\bar{D}}$	1.306			

Table 7.2.3.2: Binormal parameter estimates for girls' reading scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.459	0.010	43.925	0.000
$\bar{x}_D$	-1.288	0.039	-33.364	0.000
s_D	1.714	0.012	145.859	0.000
$\bar{x}_{\bar{D}}$	0.649			
$s_{\bar{D}}$	1.066			

Table 7.2.4.1: Binormal parameter estimates for boys' reading scores in P7

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.406	0.010	39.786	0.000
$\bar{x}_D$	-0.830	0.037	-22.164	0.000
s_D	1.620	0.010	159.005	0.000
$\bar{x}_{\bar{D}}$	0.872			
$s_{\bar{D}}$	1.045			

Table 7.2.4.2: Binormal parameter estimates for girls' reading scores in P7

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.292	0.121	2.414	0.016
$\bar{x}_D$	-0.724	0.319	-2.272	0.023
s_D	1.130	0.067	16.891	0.000
$\bar{x}_{\bar{D}}$	0.251			
$s_{\bar{D}}$	0.965			

Table 7.3.1.1: Binormal parameter estimates for boys' mathematics scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.054	0.001	73.922	0.000
$\bar{x}_D$	-1.893	0.011	-170.889	0.000
s_D	0.708	0.007	100.151	0.000
$\bar{x}_{\bar{D}}$	0.066			
$s_{\bar{D}}$	0.906			

Table 7.3.1.2: Binormal parameter estimates for girls' mathematics scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.366	0.004	102.664	0.000
$\bar{x}_D$	-0.771	0.010	-73.931	0.000
s_D	1.275	0.002	546.217	0.000
$\bar{x}_{\bar{D}}$	0.441			
$s_{\bar{D}}$	0.811			

Table 7.3.2.1: Binormal parameter estimates for boys' mathematics scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.307	0.003	89.586	0.000
$\bar{x}_D$	-0.759	0.011	-69.271	0.000
s_D	1.190	0.002	503.922	0.000
$\bar{x}_{\bar{D}}$	0.357			
$s_{\bar{D}}$	0.761			

Table 7.3.2.2: Binormal parameter estimates for girls' mathematics scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.289	0.002	155.096	0.000
$\bar{x}_D$	-0.843	0.006	-148.575	0.000
s_D	1.523	0.001	1895.414	0.000
$\bar{x}_{\bar{D}}$	0.201			
$s_{\bar{D}}$	0.881			

Table 7.3.3.1: Binormal parameter estimates for boys' mathematics scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.282	0.001	219.428	0.000
$\bar{x}_D$	-0.767	0.004	-217.325	0.000
s_D	1.424	0.001	2162.857	0.000
$\bar{x}_{\bar{D}}$	0.158			
$s_{\bar{D}}$	0.790			

Table 7.3.3.2: Binormal parameter estimates for girls' mathematics scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.790	0.002	411.824	0.000
$\bar{x}_D$	-0.191	0.003	-57.720	0.000
s_D	1.560	0.001	2275.885	0.000
$\bar{x}_{\bar{D}}$	1.225			
$s_{\bar{D}}$	0.759			

Table 7.3.4.1: Binormal parameter estimates for boys' mathematics scores in P7

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.648	0.004	183.828	0.000
$\bar{x}_D$	-0.324	0.007	-49.490	0.000
s_D	1.447	0.001	1048.998	0.000
$\bar{x}_{\bar{D}}$	0.998			
$s_{\bar{D}}$	0.859			

Table 7.3.4.2: Binormal parameter estimates for girls' mathematics scores in P7

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.191	0.011	18.121	0.000
$\bar{x}_D$	-1.476	0.086	-17.187	0.000
s_D	1.513	0.028	54.305	0.000
$\bar{x}_{\bar{D}}$	0.330			
$s_{\bar{D}}$	0.949			

Table 7.4.1.1: Binormal parameter estimates for boys' arithmetic scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.177	0.005	34.562	0.000
$\bar{x}_D$	-1.679	0.047	-35.912	0.000
s_D	1.471	0.016	93.371	0.000
$\bar{x}_{\bar{D}}$	0.135			
$s_{\bar{D}}$	0.789			

Table 7.4.1.2: Binormal parameter estimates for girls' arithmetic scores in P4

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.219	0.005	46.098	0.000
$\bar{x}_D$	-1.219	0.031	-38.707	0.000
s_D	1.720	0.006	280.849	0.000
$\bar{x}_{\bar{D}}$	0.406			
$s_{\bar{D}}$	0.956			

Table 7.4.2.1: Binormal parameter estimates for boys' arithmetic scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.249	0.005	49.753	0.000
$\bar{x}_D$	-0.986	0.024	-41.940	0.000
s_D	1.415	0.004	327.342	0.000
$\bar{x}_{\bar{D}}$	0.327			
$s_{\bar{D}}$	0.781			

Table 7.4.2.2: Binormal parameter estimates for girls' arithmetic scores in P5

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.267	0.006	42.543	0.000
$\bar{x}_D$	-0.981	0.029	-34.351	0.000
s_D	1.611	0.005	344.140	0.000
$\bar{x}_{\bar{D}}$	0.406			
$s_{\bar{D}}$	0.981			

Table 7.4.3.1: Binormal parameter estimates for boys' arithmetic scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.235	0.003	79.999	0.000
$\bar{x}_D$	-0.940	0.013	-73.046	0.000
s_D	1.698	0.003	660.562	0.000
$\bar{x}_{\bar{D}}$	0.257			
$s_{\bar{D}}$	0.810			

Table 7.4.3.2: Binormal parameter estimates for girls' arithmetic scores in P6

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.237	0.006	37.046	0.000
$\bar{x}_D$	-0.953	0.028	-34.508	0.000
s_D	2.060	0.004	473.907	0.000
$\bar{x}_{\bar{D}}$	0.311			
$s_{\bar{D}}$	1.289			

Table 7.4.4.1: Binormal parameter estimates for boys' arithmetic scores in P7

Variable	Value	Std. Error	t-ratio	Prob(t)
ρ	0.283	0.006	47.025	0.000
$\bar{x}_D$	-0.765	0.021	-36.098	0.000
s_D	1.630	0.002	693.972	0.000
$\bar{x}_{\bar{D}}$	0.311			
$s_{\bar{D}}$	1.289			

Table 7.4.4.2: Binormal parameter estimates for girls' arithmetic scores in P7

5. The Variation Explained by the Models

The coefficient of multiple determination (R^2) describes the proportion of variation in data explained by the regression model. On its own the normal model explains the bulk of the variation in all of the circumstances here described. Another way to investigate the improvement in model fit is to consider the additional variance explained by the binormal model over and above that explained by the normal distribution model. This is similar to the analysis performed in section 3 of this chapter, but is more readily interpreted. It also allows an evaluation of the magnitude of any residual unexplained variation. The results from this analysis are given in tables 8.1.1 to 8.4.2.

The partitioning of explained variance reported in tables 8.1.1 to 8.4.2 reveals some trends, but there is no entirely consistent pattern. Perhaps the most consistent trend concerned the mental arithmetic assessments. These showed the greatest amount of additional explained variance by the binormal model, together with the least amount of unexplained variance.

Not surprisingly the girls' picture vocabulary results in P7, for which the binormal model was rejected in the previous section, showed a very high proportion of variance explained by the normal model. However the girls' P4 mathematics scores and the P4 reading scores of both boys and girls all showed a higher proportion of variance explained by the normal model, even though the binormal model provided a statistically better fit. Interestingly the girls' P7 picture vocabulary results showed a relatively high proportion of unexplained variance after fitting the binormal model, and this suggests that the curve fitting methodology may have arrived at an incorrect solution in this case.

The lowest amount of additional variation explained by the binormal model was 0.038% in the case of boys' reading results in P4, and this still gave a significantly better fit than the normal model. For several assessments, most notably the P7 reading results for both boys and girls, the amount of unexplained variance after fitting the binormal model was of a similar magnitude. This

suggests that in some circumstances there may be scope to fit a model to the data that is even more complicated than the binormal distribution.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.99454	0.99974	0.52%	0.03%
P5	0.99570	0.99980	0.41%	0.02%
P6	0.99472	0.99987	0.52%	0.01%
P7	0.99586	0.99990	0.40%	0.01%

Table 8.1.1: Comparison of the variation explained by the normal and binormal models applied to boys' picture vocabulary scores.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.99399	0.99966	0.57%	0.03%
P5	0.99069	0.99991	0.92%	0.01%
P6	0.99478	0.99956	0.48%	0.04%
P7	0.99771	0.99944	0.17%	0.06%

Table 8.1.2: Comparison of the variation explained by the normal and binormal models applied to girls' picture vocabulary scores.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.99956	0.99994	0.04%	0.01%
P5	0.99478	0.99990	0.51%	0.01%
P6	0.98976	0.99980	1.00%	0.02%
P7	0.98638	0.99926	1.29%	0.07%

Table 8.2.1: Comparison of the variation explained by the normal and binormal models applied to boys' reading scores.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.99931	0.99986	0.06%	0.01%
P5	0.99587	0.99985	0.40%	0.01%
P6	0.99436	0.99979	0.54%	0.02%
P7	0.98847	0.99953	1.11%	0.05%

Table 8.2.2: Comparison of the variation explained by the normal and binormal models applied to girls' reading scores.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.99868	0.99981	0.11%	0.02%
P5	0.98913	0.99991	1.08%	0.01%
P6	0.99001	0.99991	0.99%	0.01%
P7	0.99515	0.99966	0.45%	0.03%

Table 8.3.1: Comparison of the variation explained by the normal and binormal models applied to boys' mathematics scores.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.99828	0.99995	0.17%	0.01%
P5	0.98955	0.99993	1.04%	0.01%
P6	0.98887	0.99995	1.11%	0.01%
P7	0.99372	0.99979	0.61%	0.02%

Table 8.3.2: Comparison of the variation explained by the normal and binormal models applied to girls' mathematics scores.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.98483	0.99988	1.50%	0.01%
P5	0.98234	0.99982	1.75%	0.02%
P6	0.98887	0.99984	1.10%	0.02%
P7	0.99495	0.99986	0.49%	0.01%

Table 8.4.1: Comparison of the variation explained by the normal and binormal models applied to boys' arithmetic scores.

Year Group	Coefficient of Multiple Determination (R^2)		Additional Variation Explained	Unexplained Variation
	Normal Model	Binormal Model		
P4	0.97119	0.99986	2.87%	0.01%
P5	0.98111	0.99979	1.87%	0.02%
P6	0.97877	0.99979	2.10%	0.02%
P7	0.97464	0.99989	2.53%	0.01%

Table 8.4.2: Comparison of the variation explained by the normal and binormal models applied to girls' arithmetic scores.

6. Summary

An examination of the output produced by the DataFit software revealed that, with one exception, the binormal distribution model provided a statistically better fit than the normal distribution model. The exception was found in the case of the picture vocabulary scores for the oldest cohort of girls (P7). In this case it was accepted that the normal distribution represented a more appropriate model, although the possibility that the software had settled upon an incorrect solution was also considered. It was also argued that in some cases, such as those of the reading results for the P7 cohort, the distribution of scores may be even more complicated than that allowed for by the binormal model.

Chapter 6: A Visual Examination of Model Fits

1. Introduction

In chapter 5 it was established that in most instances the binormal model provided a better fit for the distribution of assessment scores than did the normal model. Whilst this is encouraging, a test of statistical significance on its own does not guarantee the validity of a particular model. The model must also make sense within a theoretical framework. The first stage in establishing this is to consider the face validity of the model fits.

In this chapter three graphs are presented for each dataset. Figure 2 presents the pdf curve for the normal distribution model against a probability histogram of the actual age-ability difference scores. The parameter values used to produce the normal curve are given in table 2. Figure 3 presents the pdf curve for the binormal distribution model against the same probability histogram. The parameter values used to produce the binormal curve are given in table 7. Figure 3 uses the same parameter values to present the normal pdf curve of each of the subpopulations that make up the binormal model fit. In these subpopulation plots the distribution of the low attaining subpopulation is represented by a blue line and that of the higher attaining subpopulation by a pink line.

2. The Production of Probability Histograms

The histograms presented in figure 2 and figure 3 take the usual frequency of observations in each category and rescale them as a probability. This procedure places the histogram on the same scale as the pdf curve, thus allowing a direct visual inspection of the model fit.

For the production of each histogram a category width (w) of 0.25 years was chosen. The age-ability difference scores were rounded to the nearest 0.25 years. The scores were then aggregated by this rounded number and the frequency of observations (f) in each category determined. The height of each bar in the histogram was then determined using the following expression in which N refers to the total of all the observations.

$$P(X) = \frac{f}{wN}$$

This expression simply takes the proportion of observations in each category and then makes a correction for the category width.

Figure 2.1.1.1: Normal Model Plot for Boys' Picture Vocabulary Results in P4

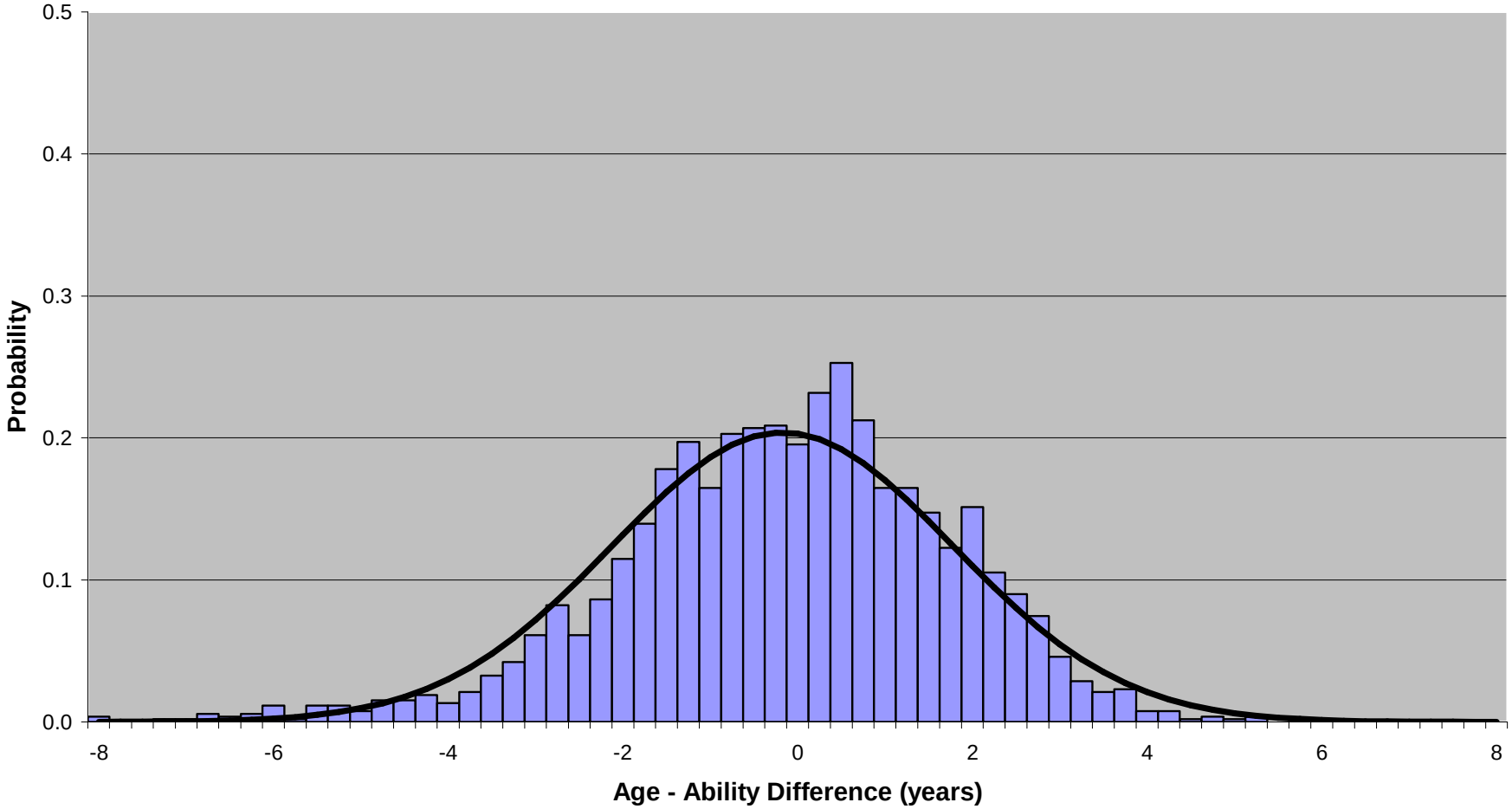


Figure 2.1.1.2: Normal Model Plot for Girls' Picture Vocabulary Results in P4

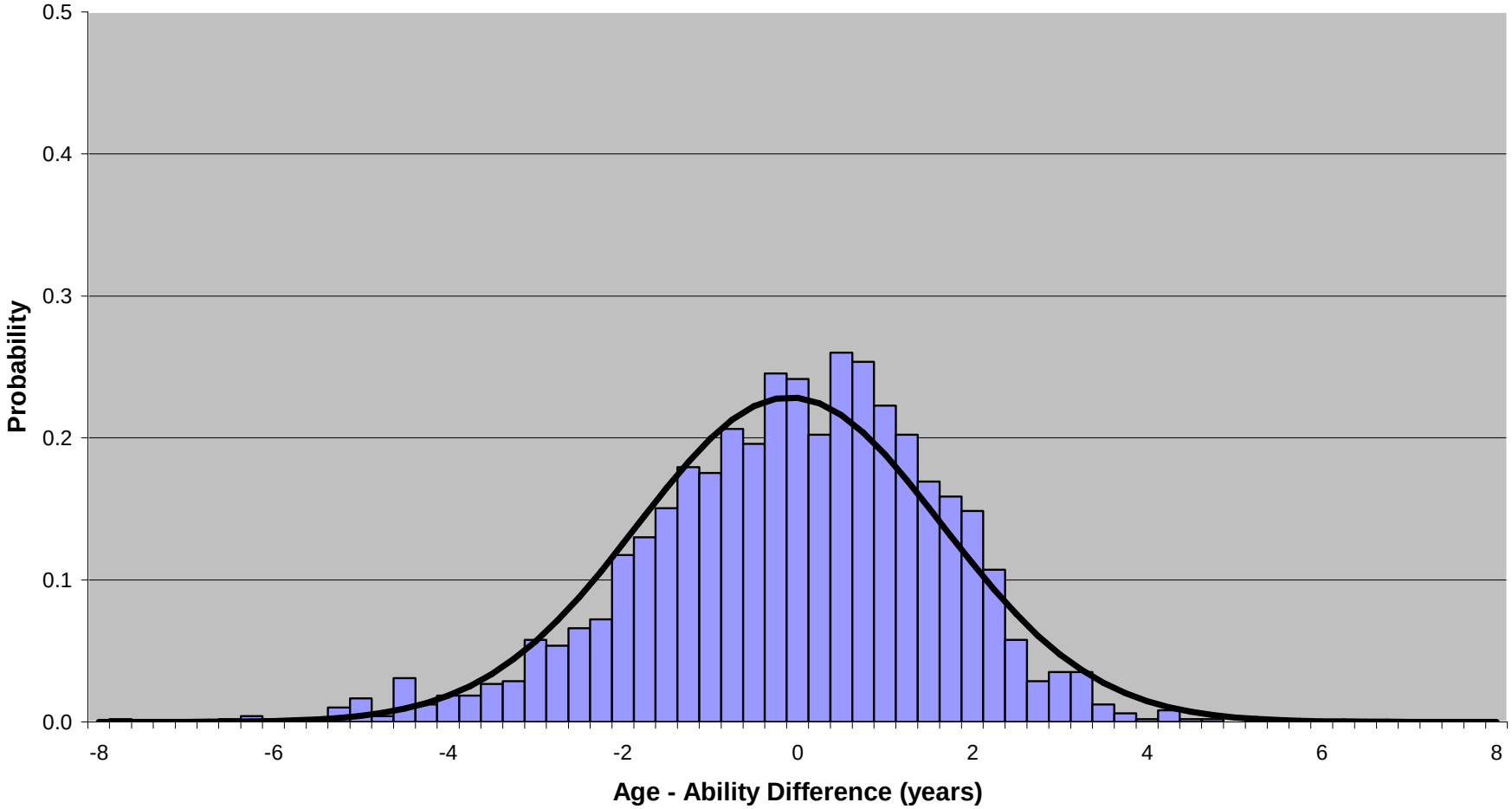


Figure 2.1.2.1: Normal Model Plot for Boys' Picture Vocabulary Results in P5

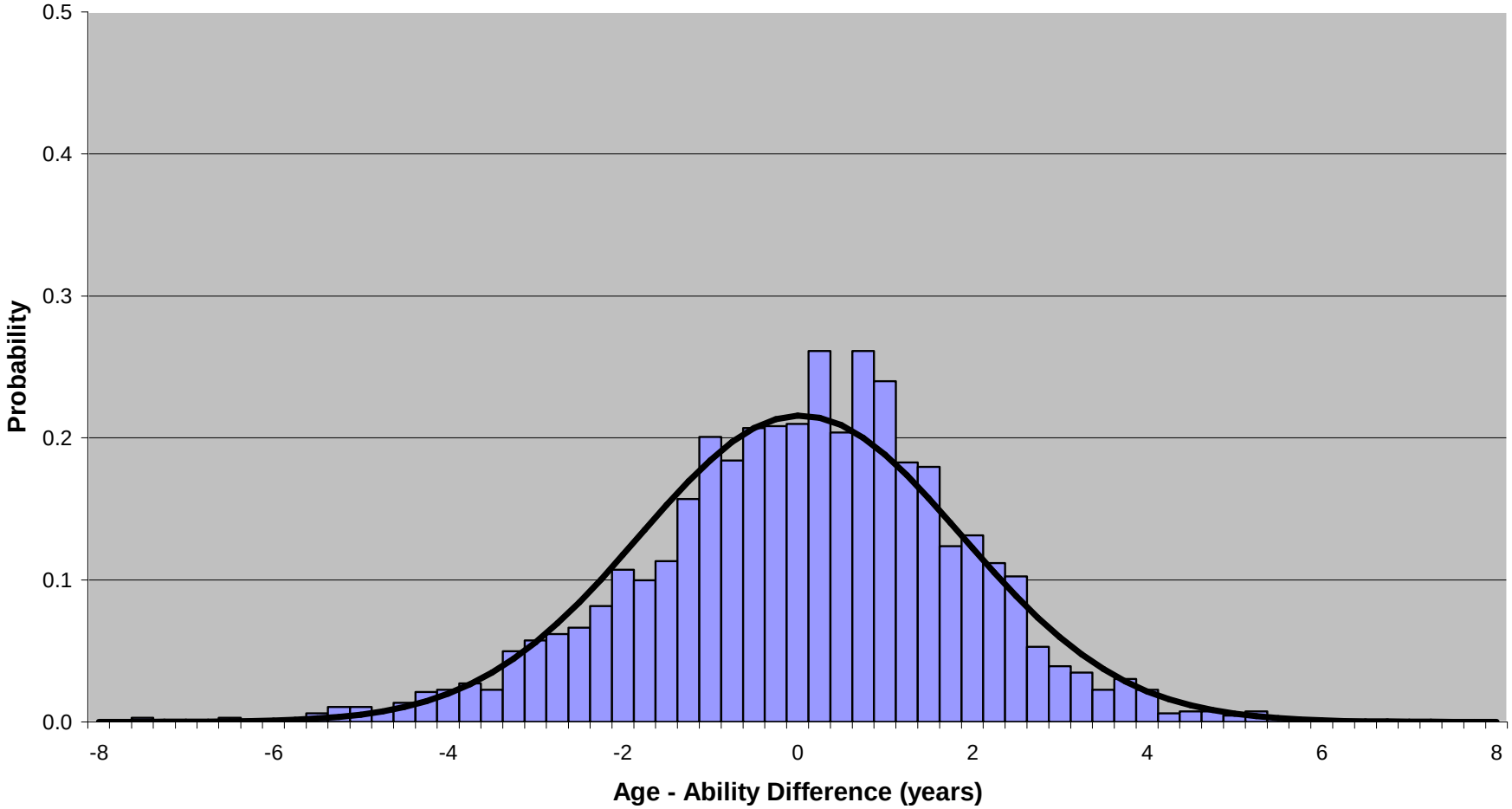


Figure 2.1.2.2: Normal Model Plot for Girls' Picture Vocabulary Results in P5

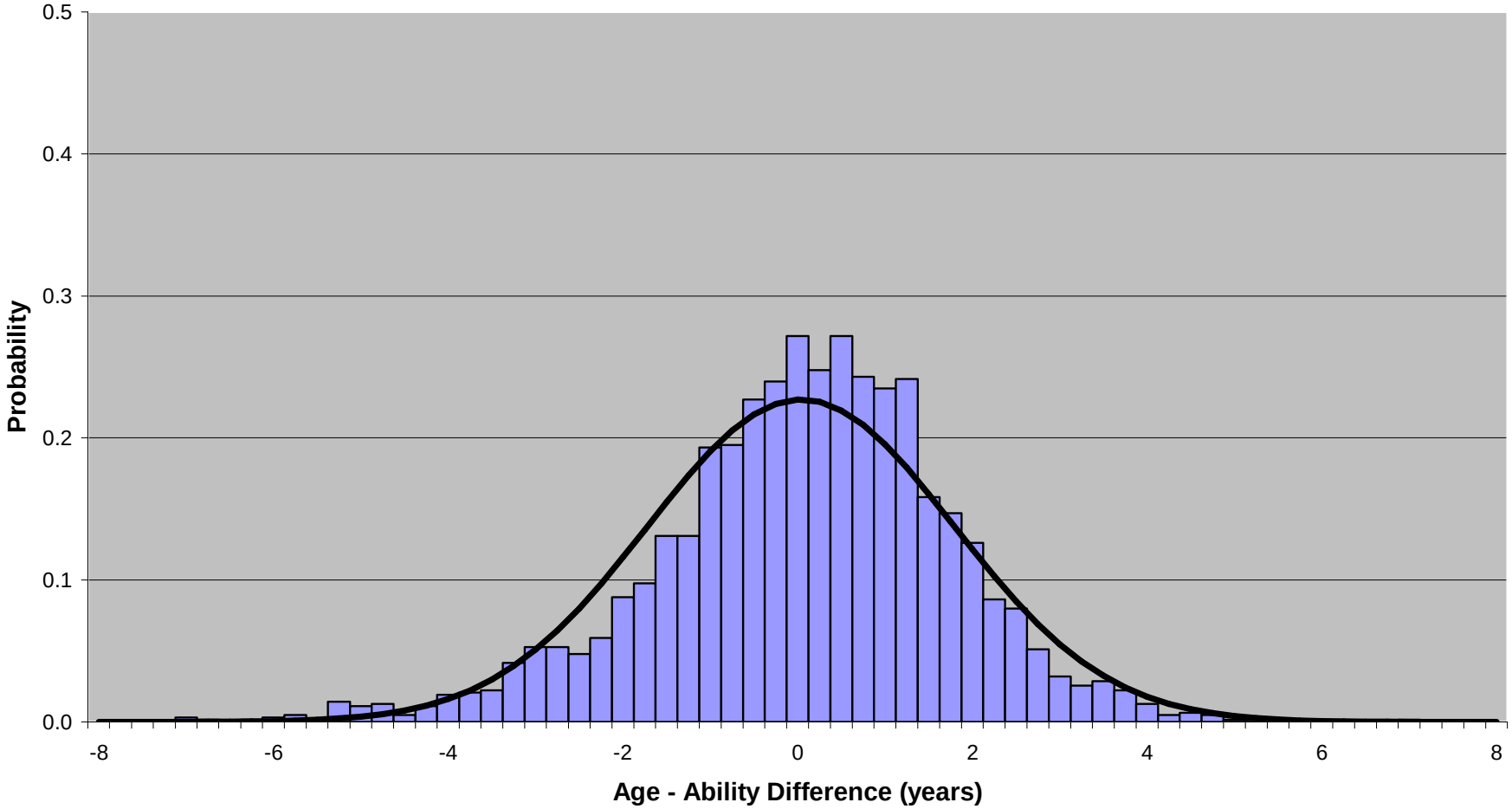


Figure 2.1.3.1: Normal Model Plot for Boys' Picture Vocabulary Results in P6

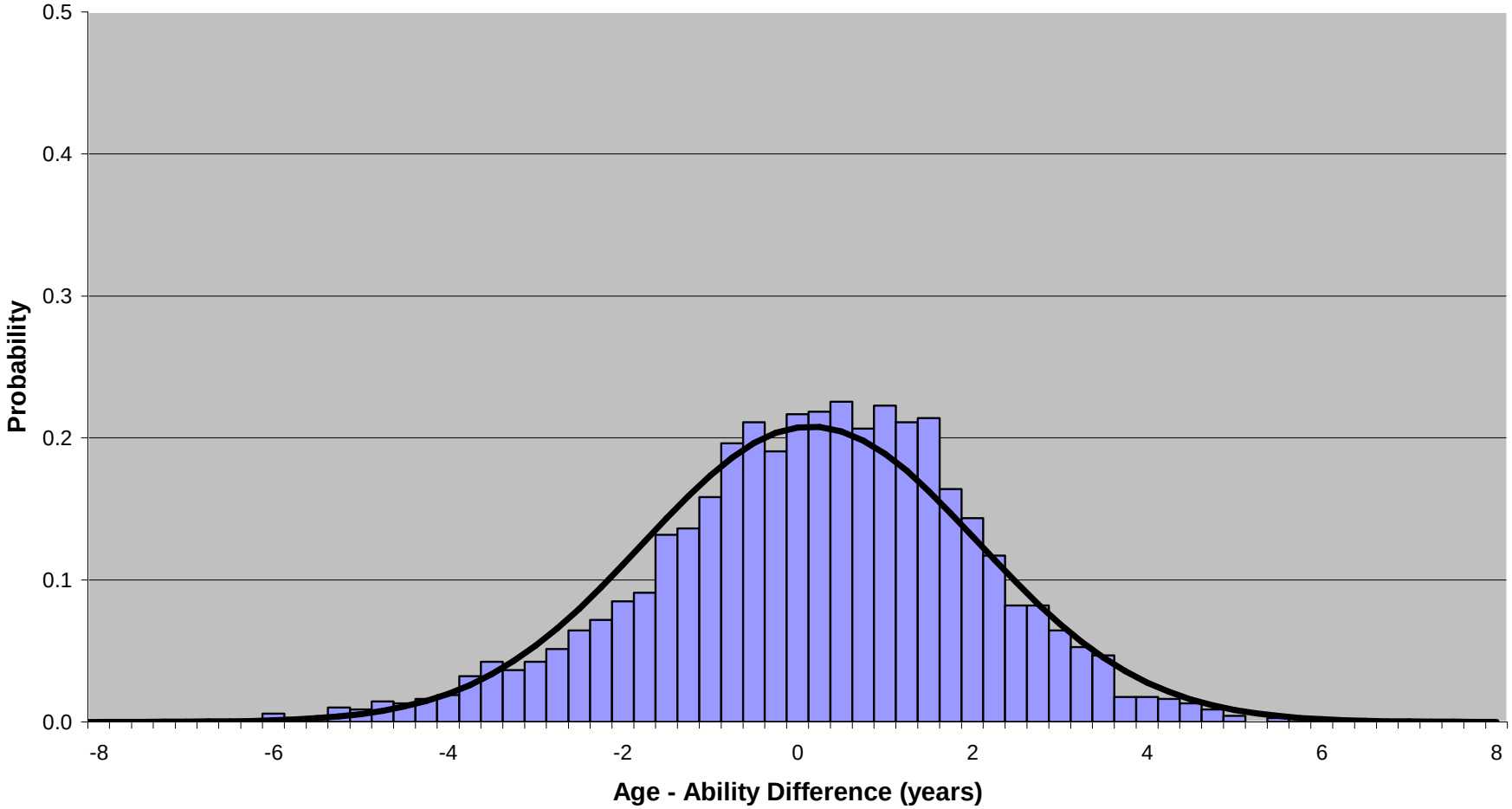


Figure 2.1.3.2: Normal Model Plot for Girls' Picture Vocabulary Results in P6

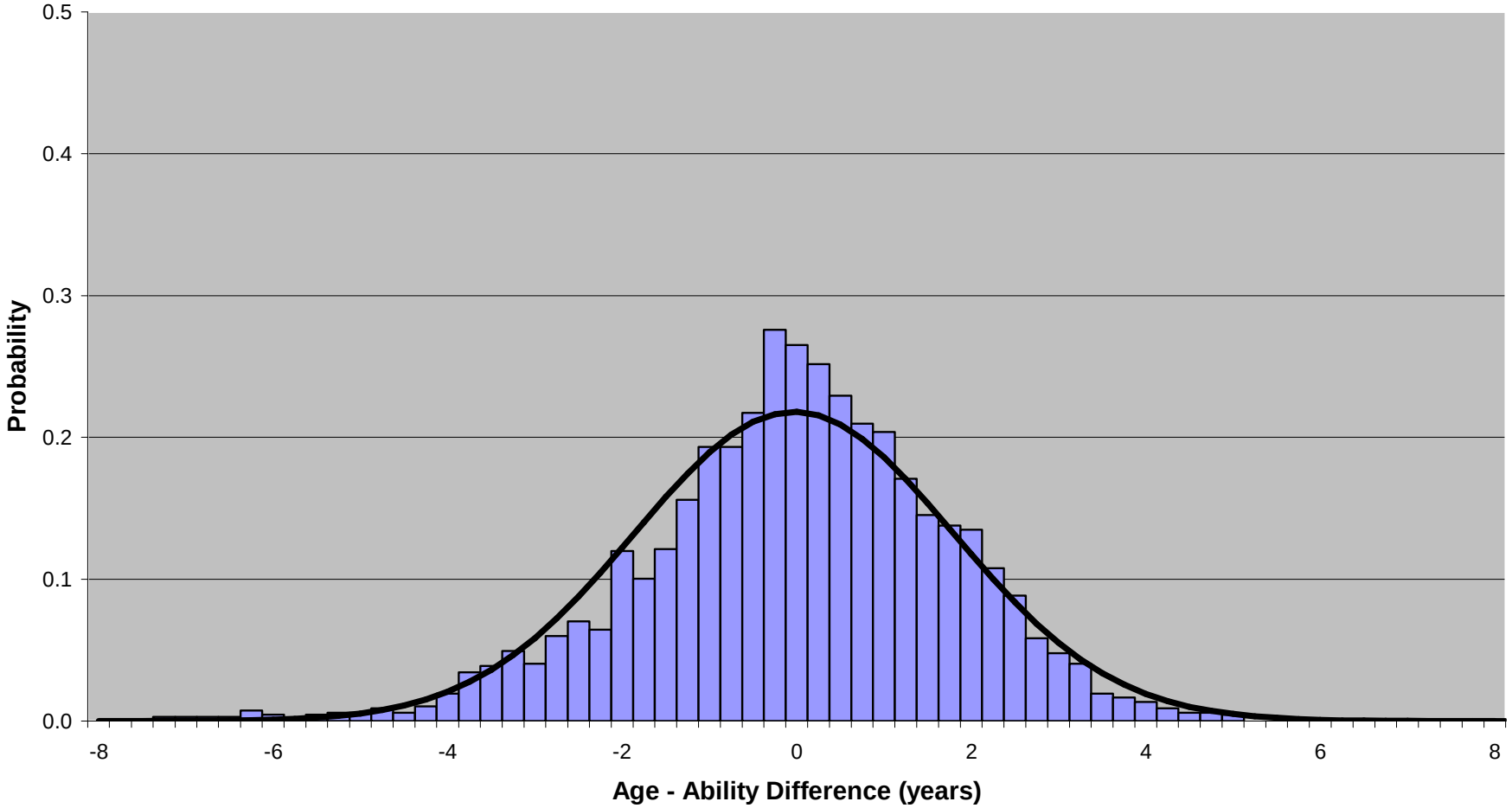


Figure 2.1.4.1: Normal Model Plot for Boys' Picture Vocabulary Results in P7

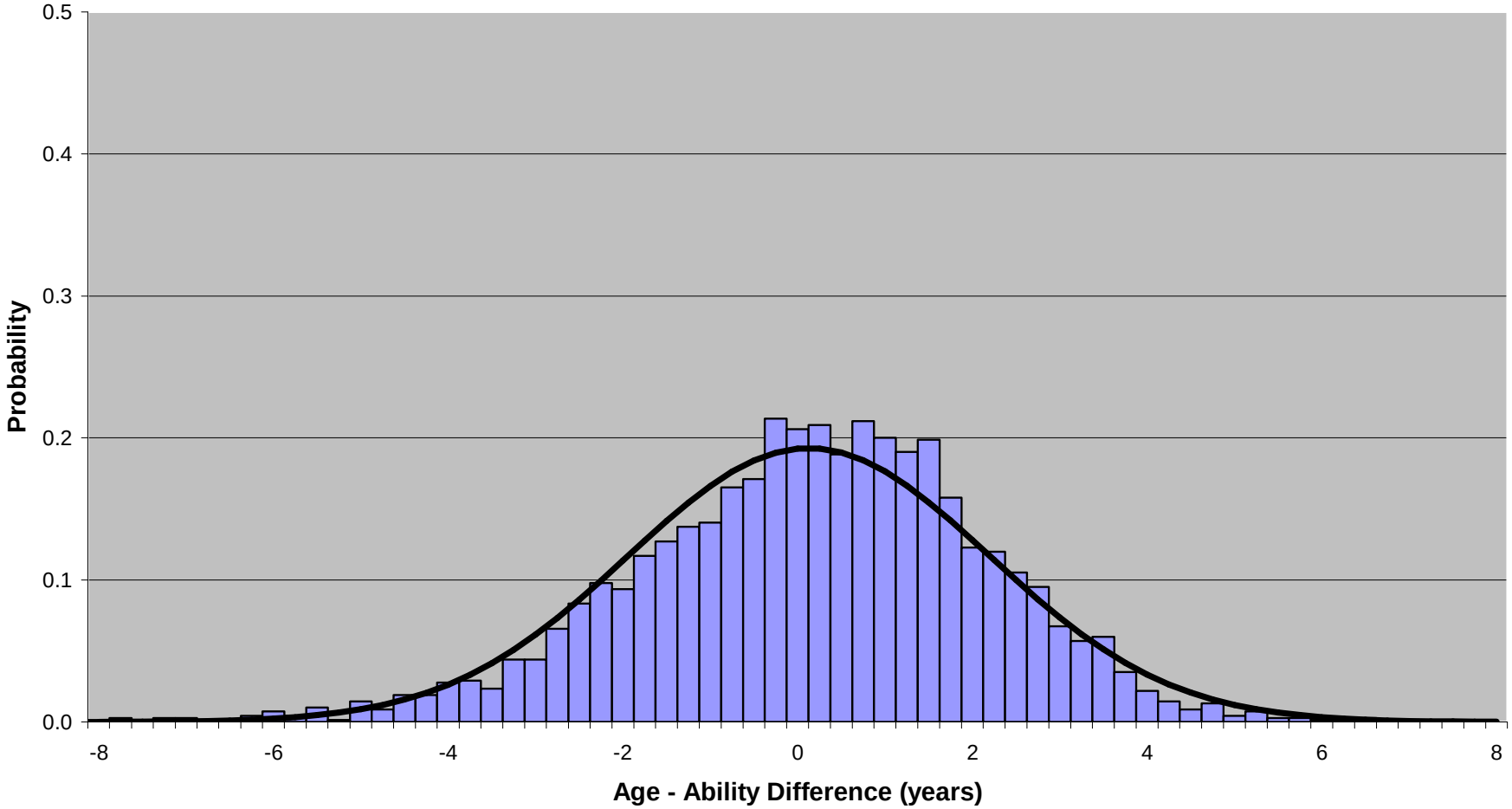


Figure 2.1.4.2: Normal Model Plot for Girls' Picture Vocabulary Results in P7

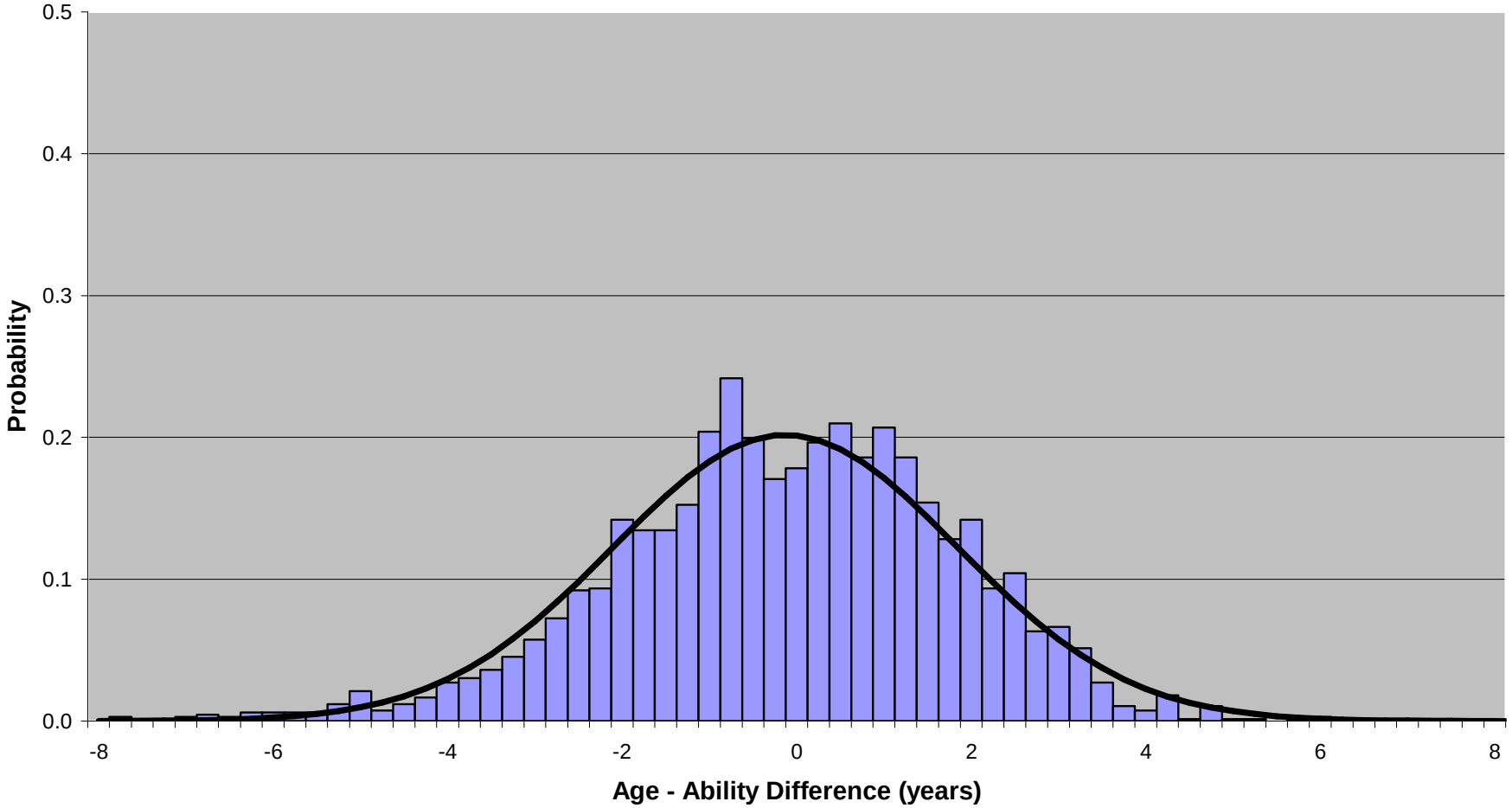


Figure 2.2.1.1: Normal Model Plot for Boys' Reading Results in P4

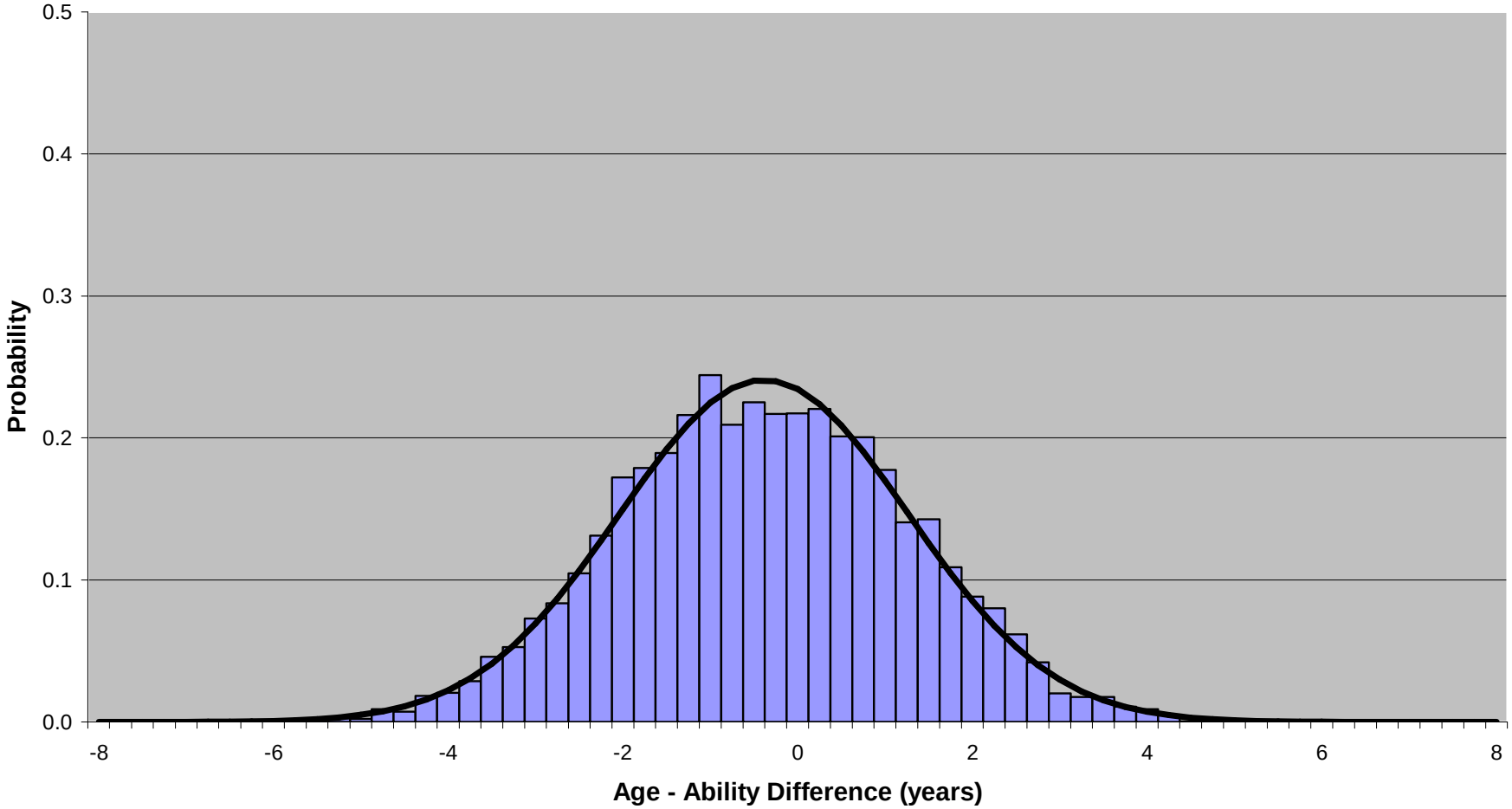


Figure 2.2.1.2: Normal Model Plot for Girls' Reading Results in P4

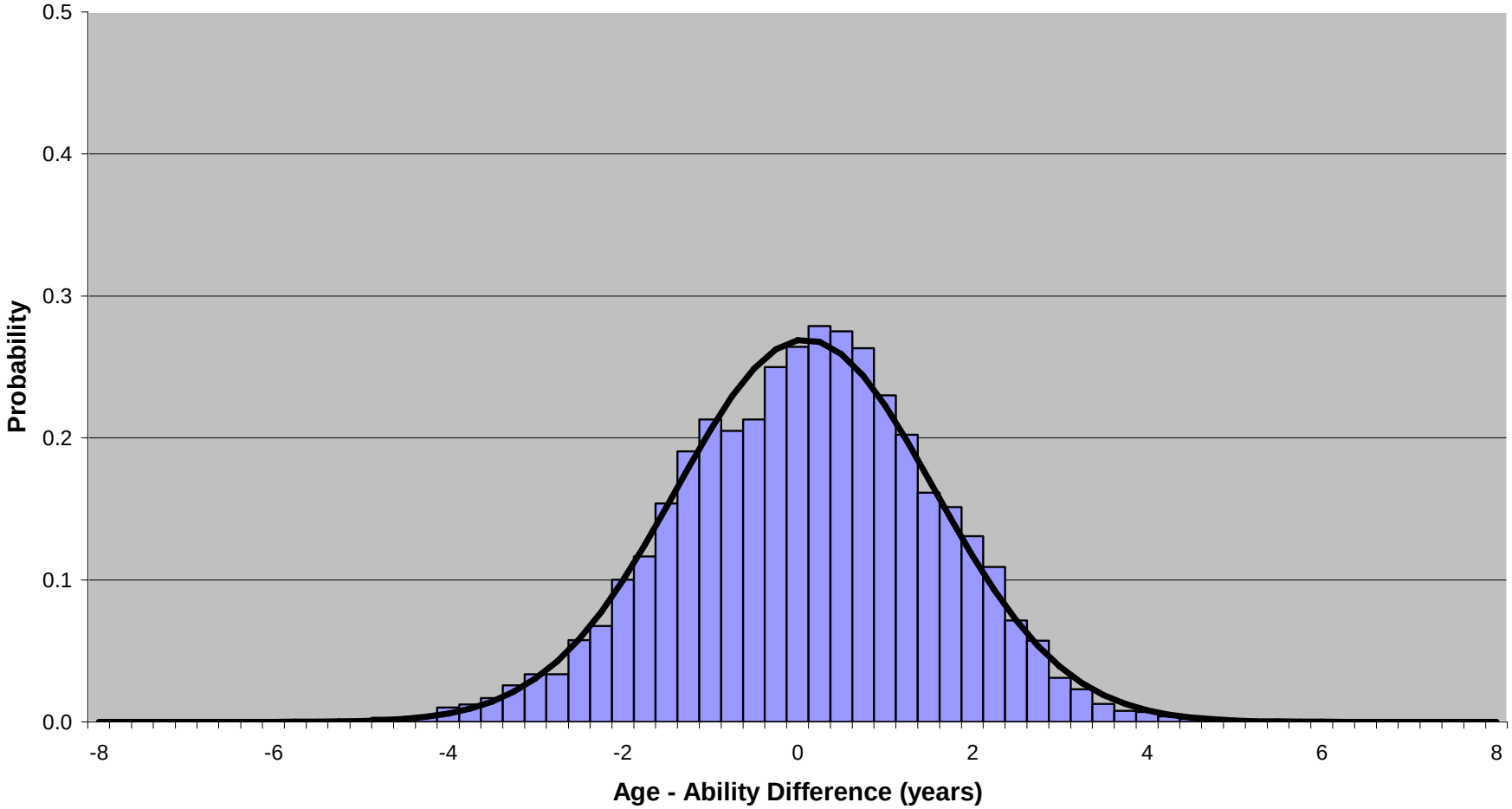


Figure 2.2.2.1: Normal Model Plot for Boys' Reading Results in P5

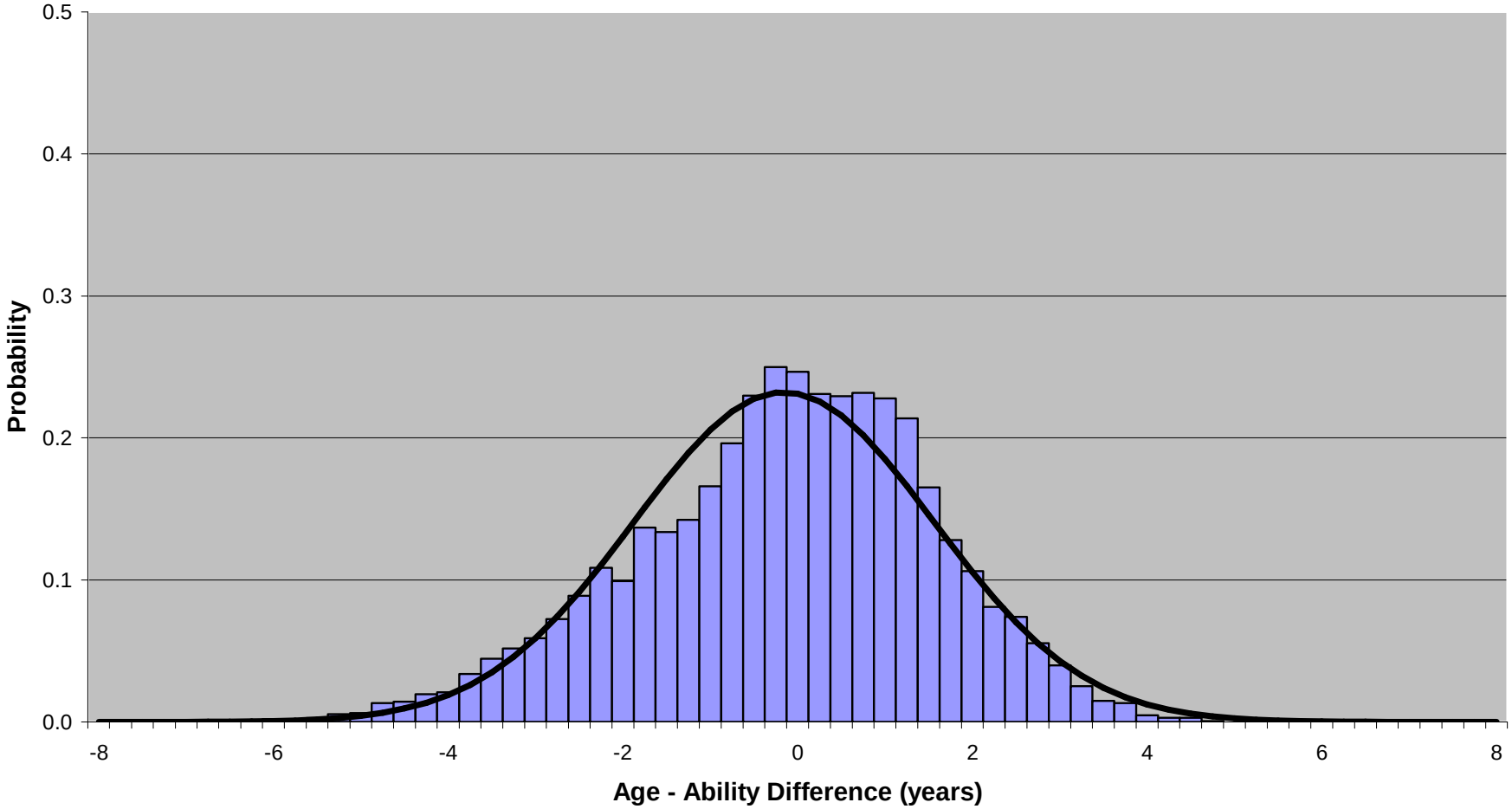


Figure 2.2.2.2: Normal Model Plot for Girls' Reading Results in P5

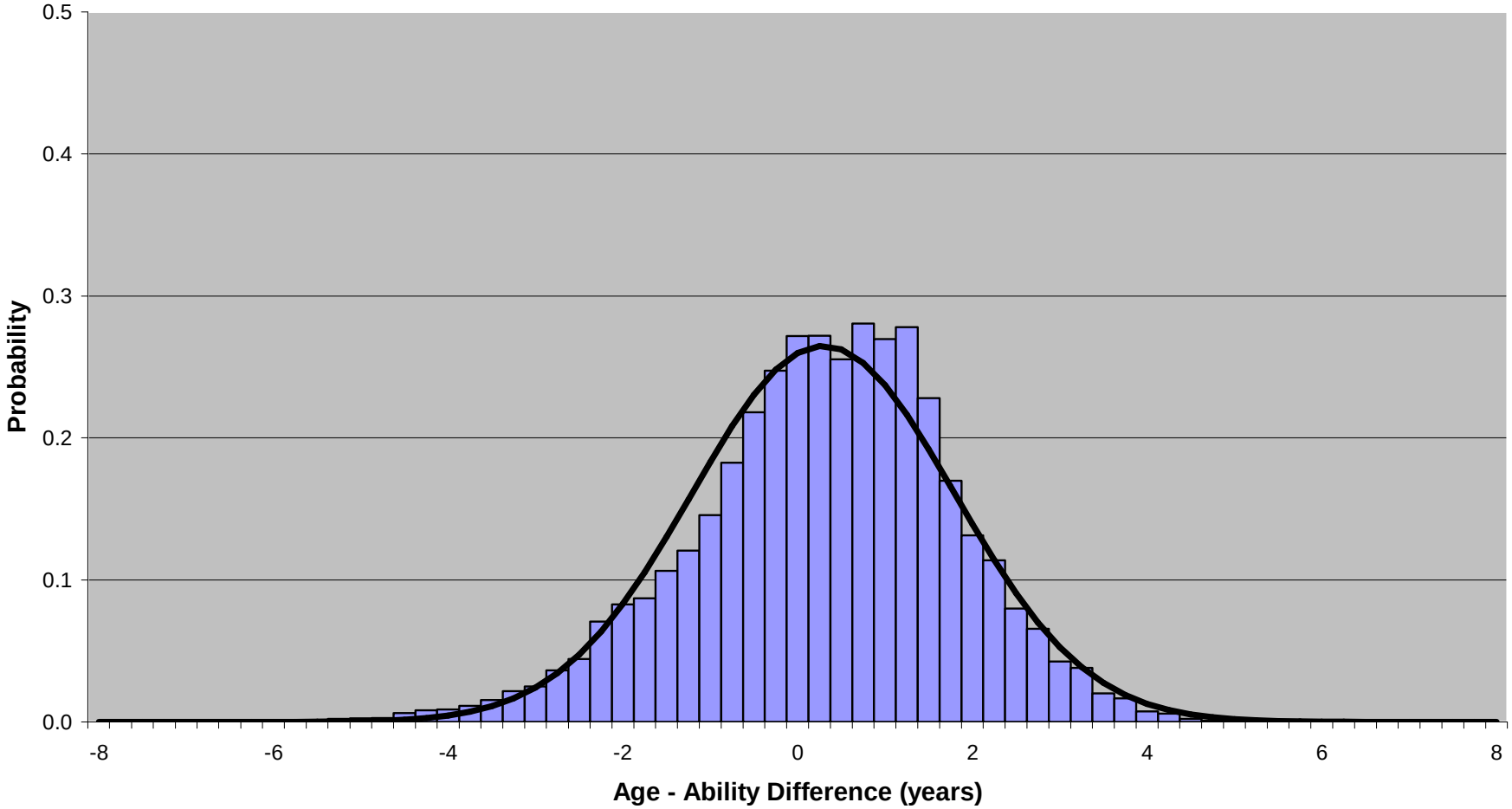


Figure 2.2.3.1: Normal Model Plot for Boys' Reading Results in P6

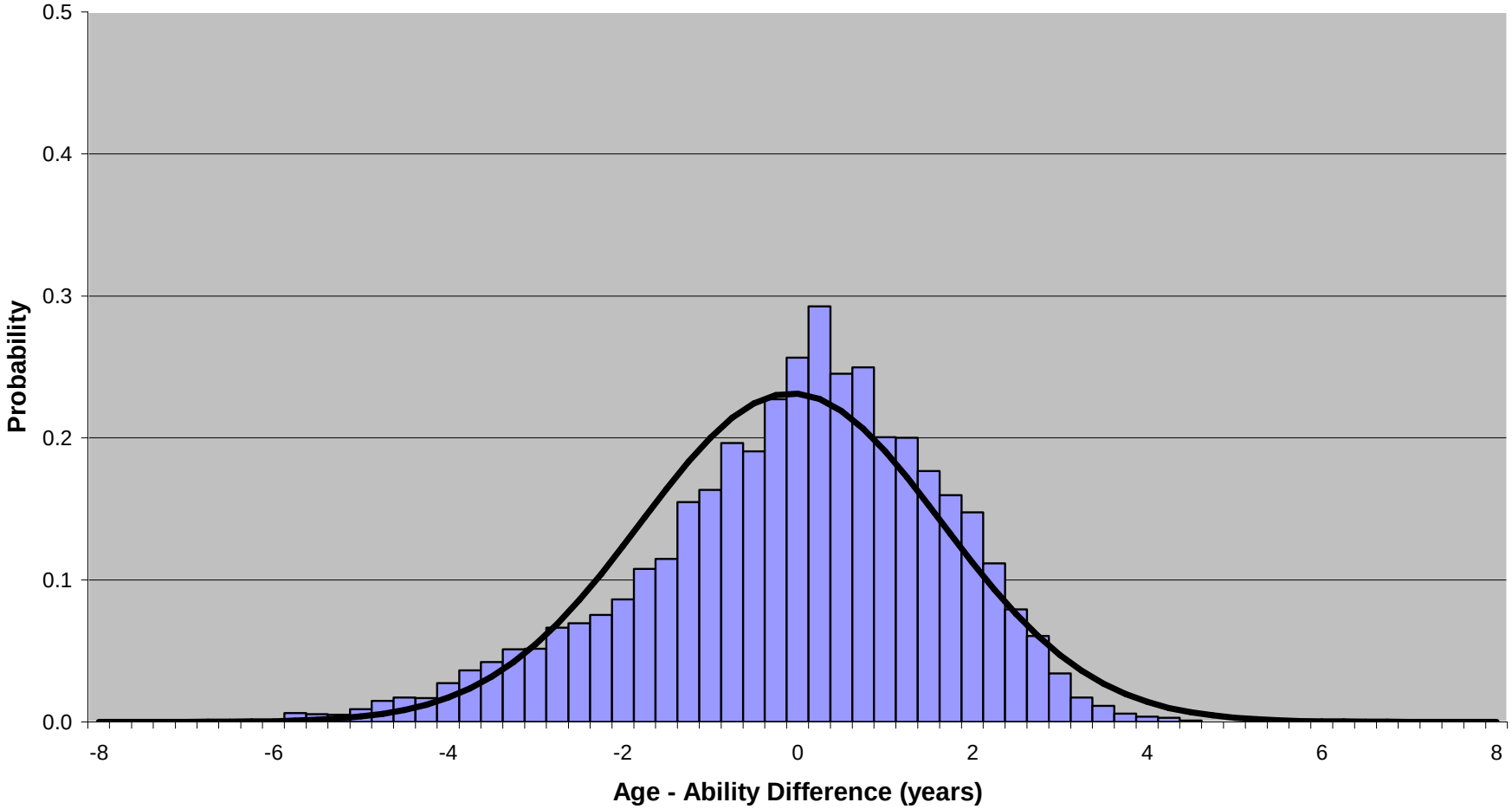


Figure 2.2.3.2: Normal Model Plot for Girls' Reading Results in P6

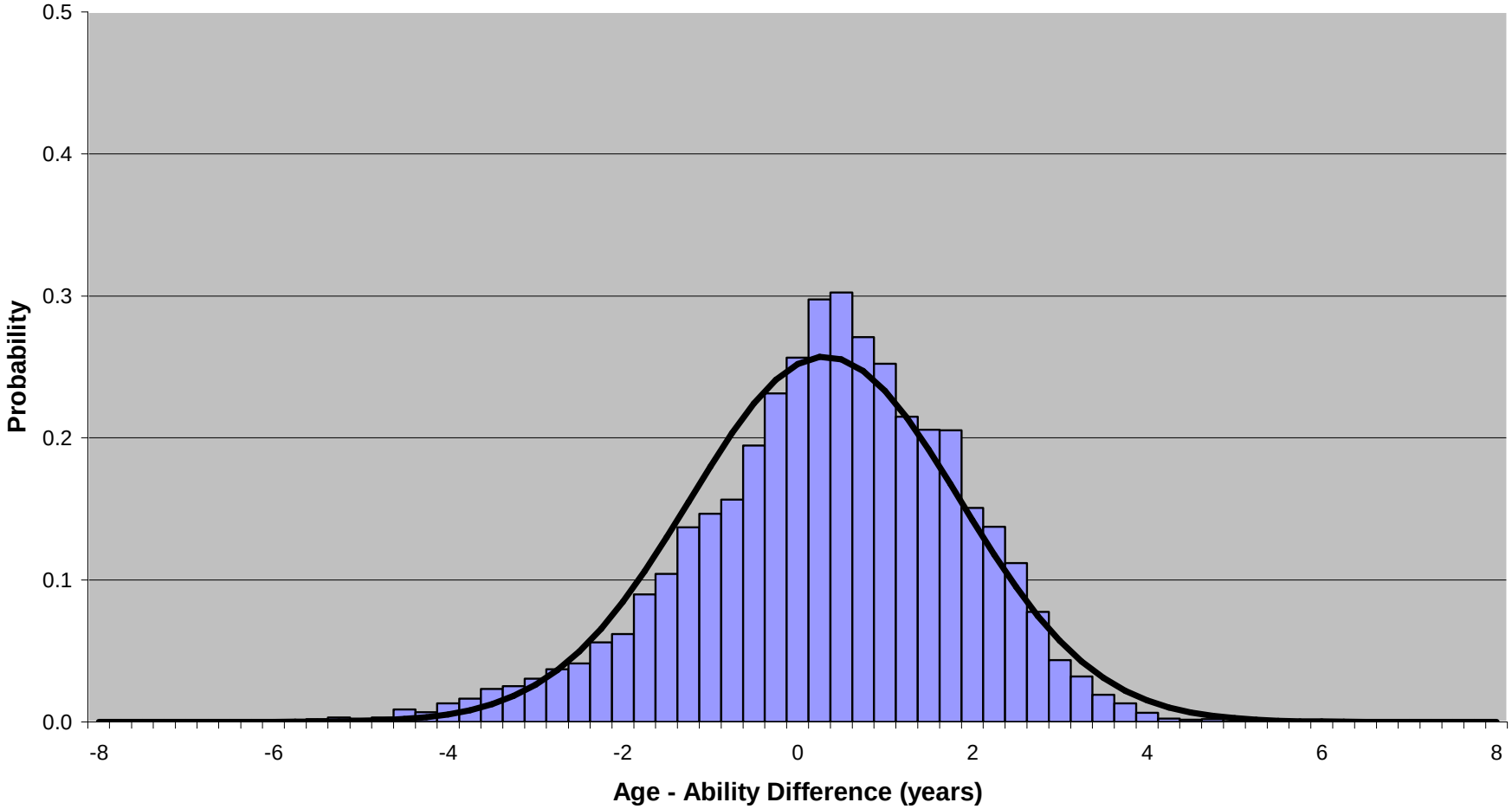


Figure 2.2.4.1: Normal Model Plot for Boys' Reading Results in P7

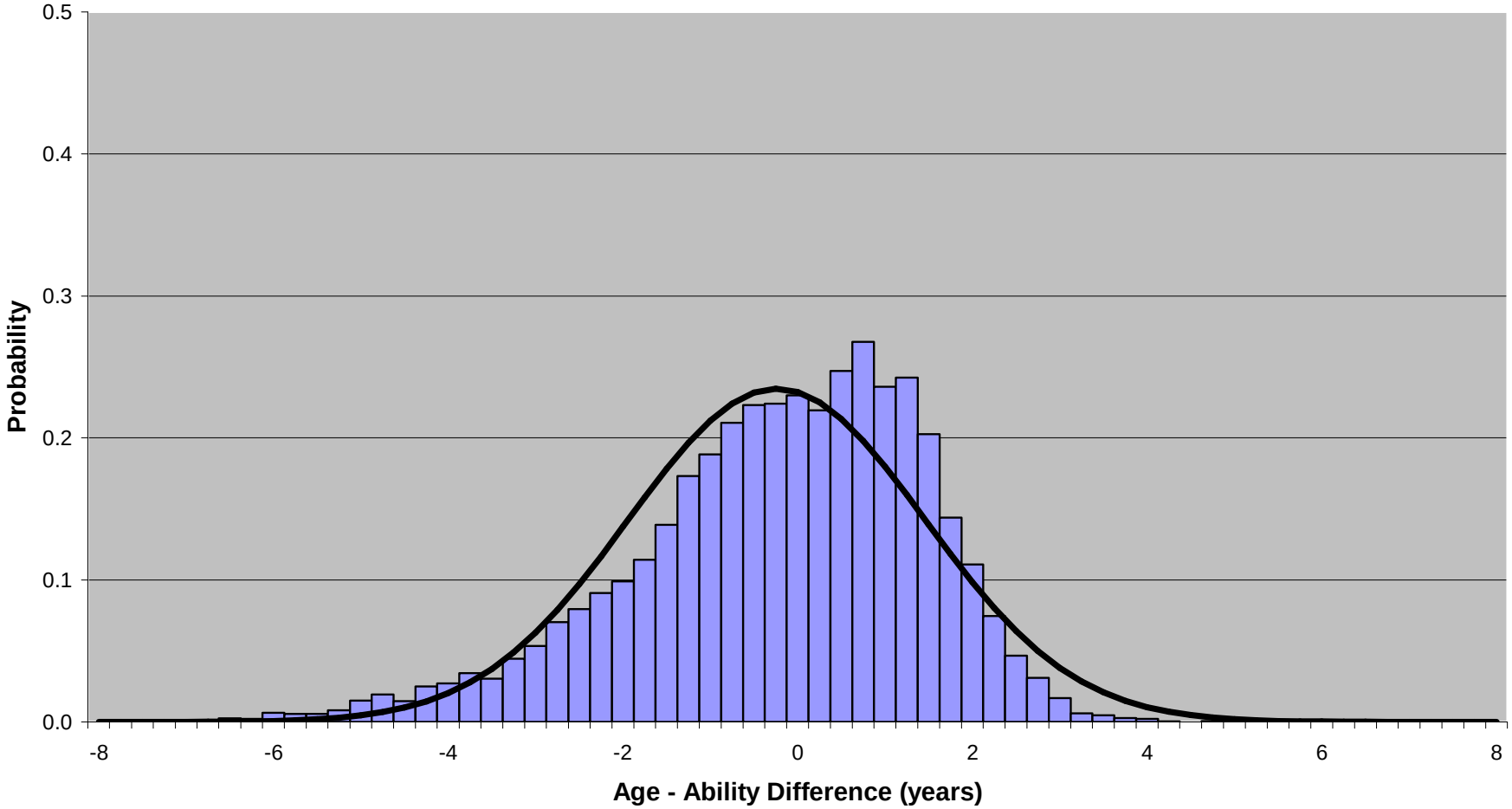


Figure 2.2.4.2: Normal Model Plot for Girls' Reading Results in P7

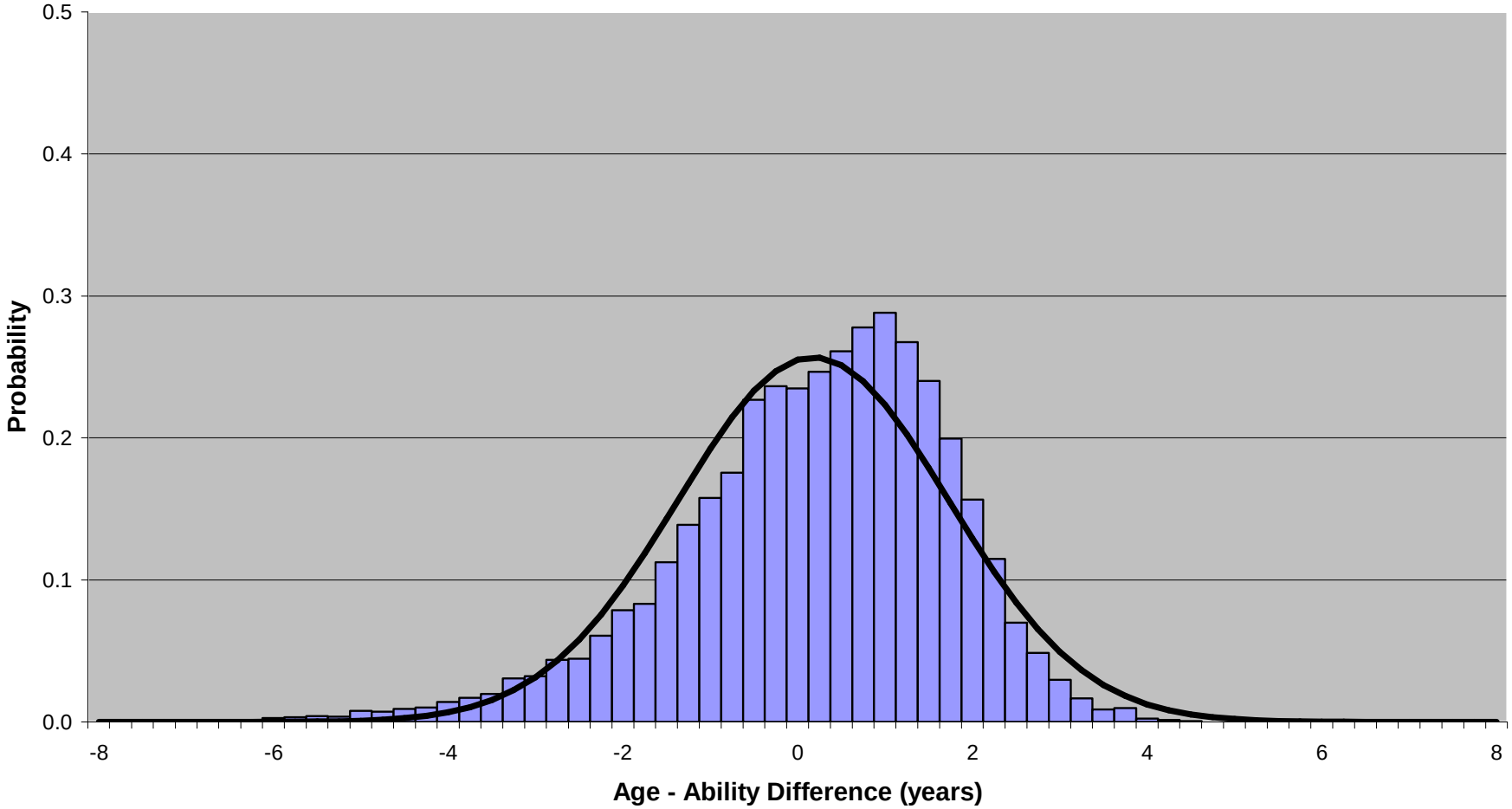


Figure 2.3.1.1: Normal Model Plot for Boys' Mathematics Results in P4

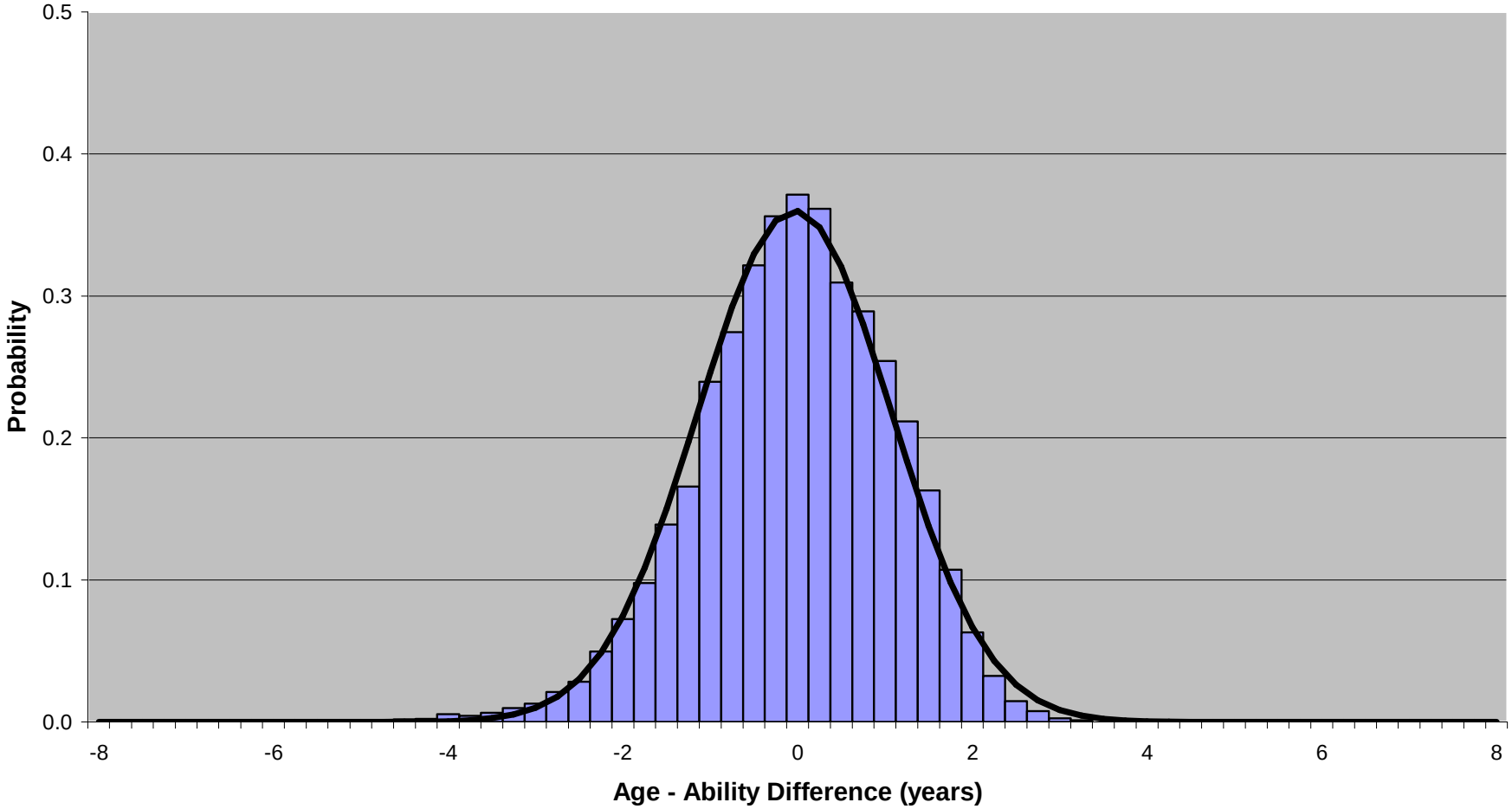


Figure 2.3.1.2: Normal Model Plot for Girls' Mathematics Results in P4

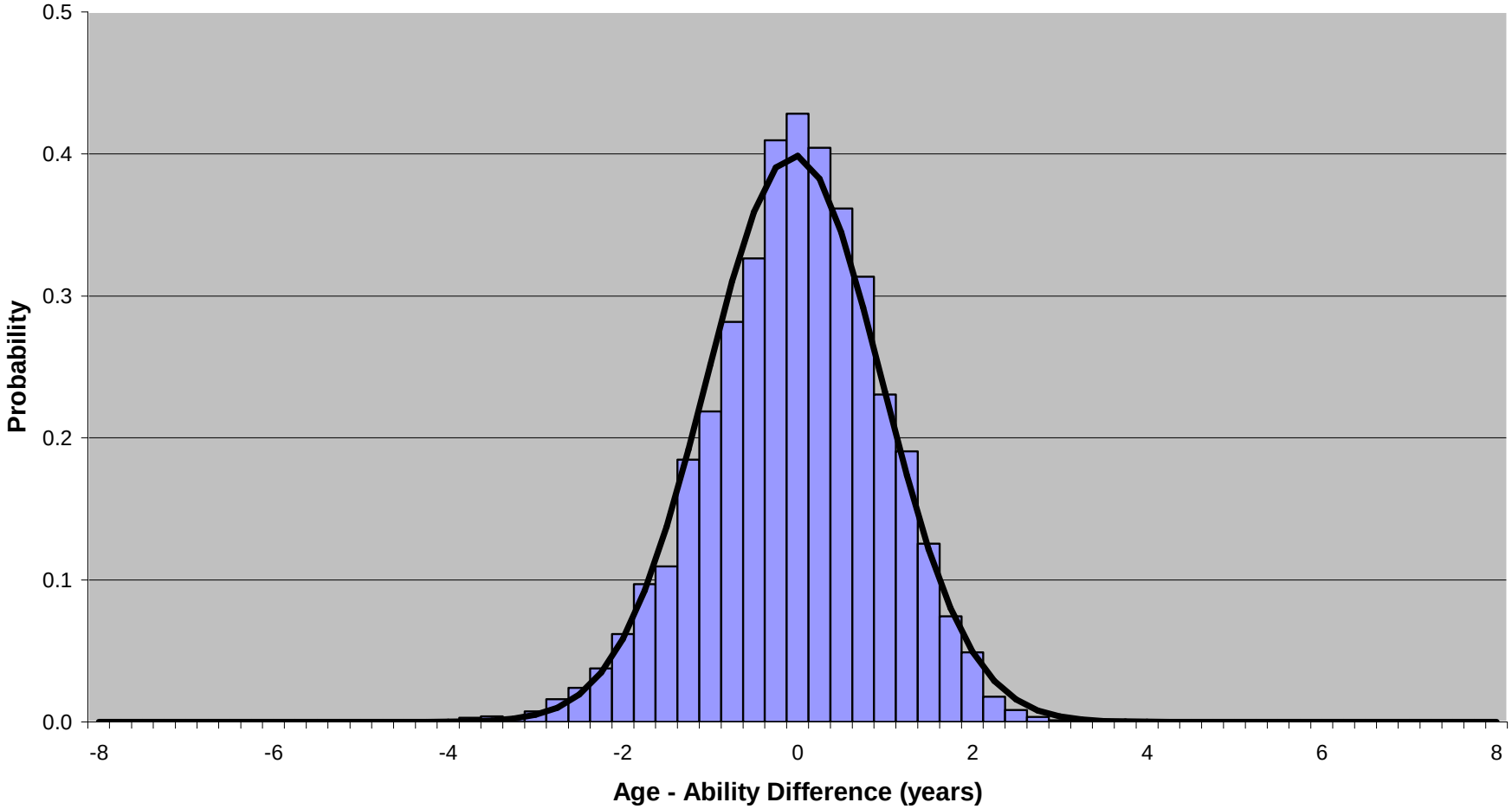


Figure 2.3.2.1: Normal Model Plot for Boys' Mathematics Results in P5

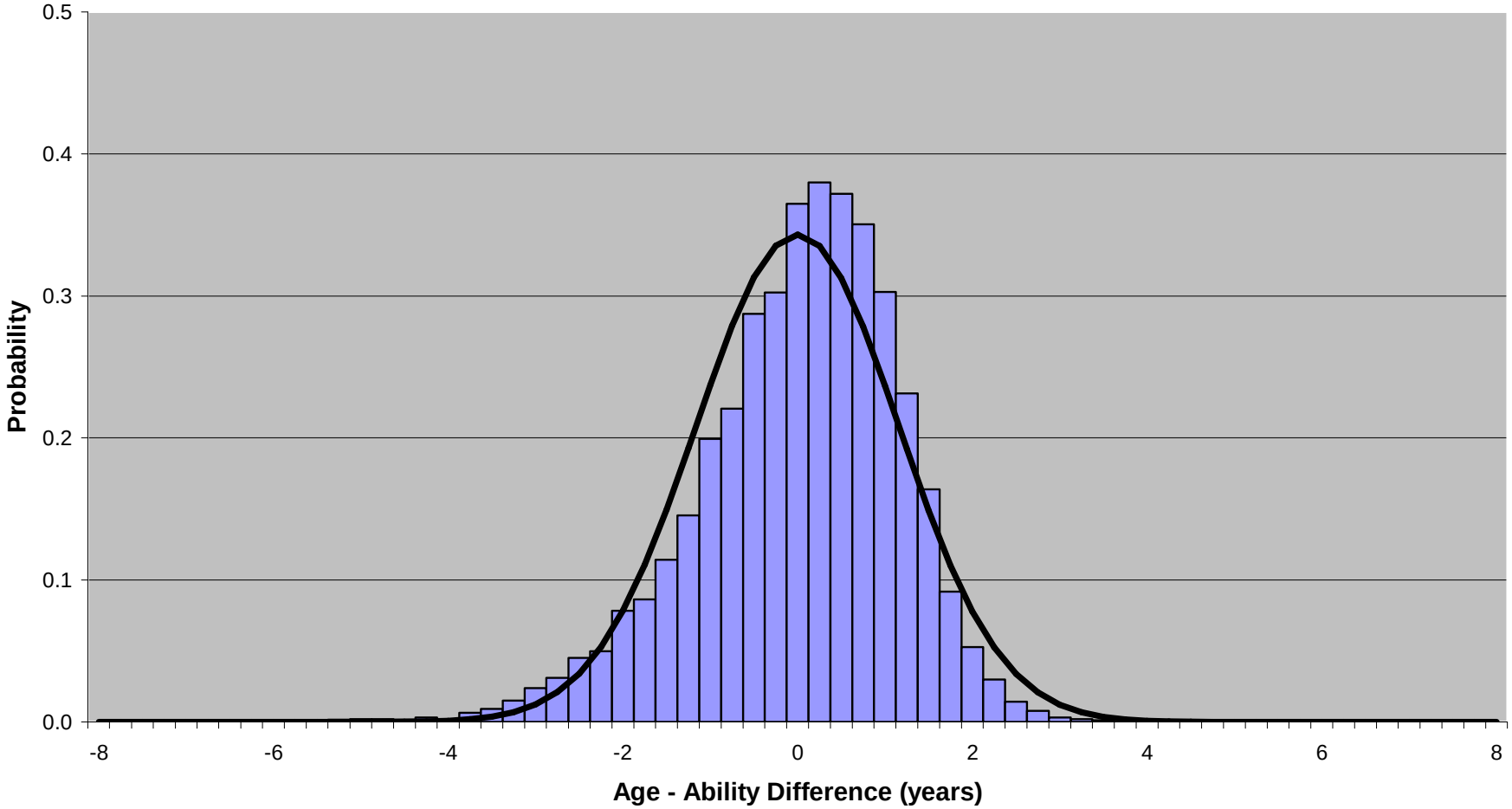


Figure 2.3.2.2: Normal Model Plot for Girls' Mathematics Results in P5

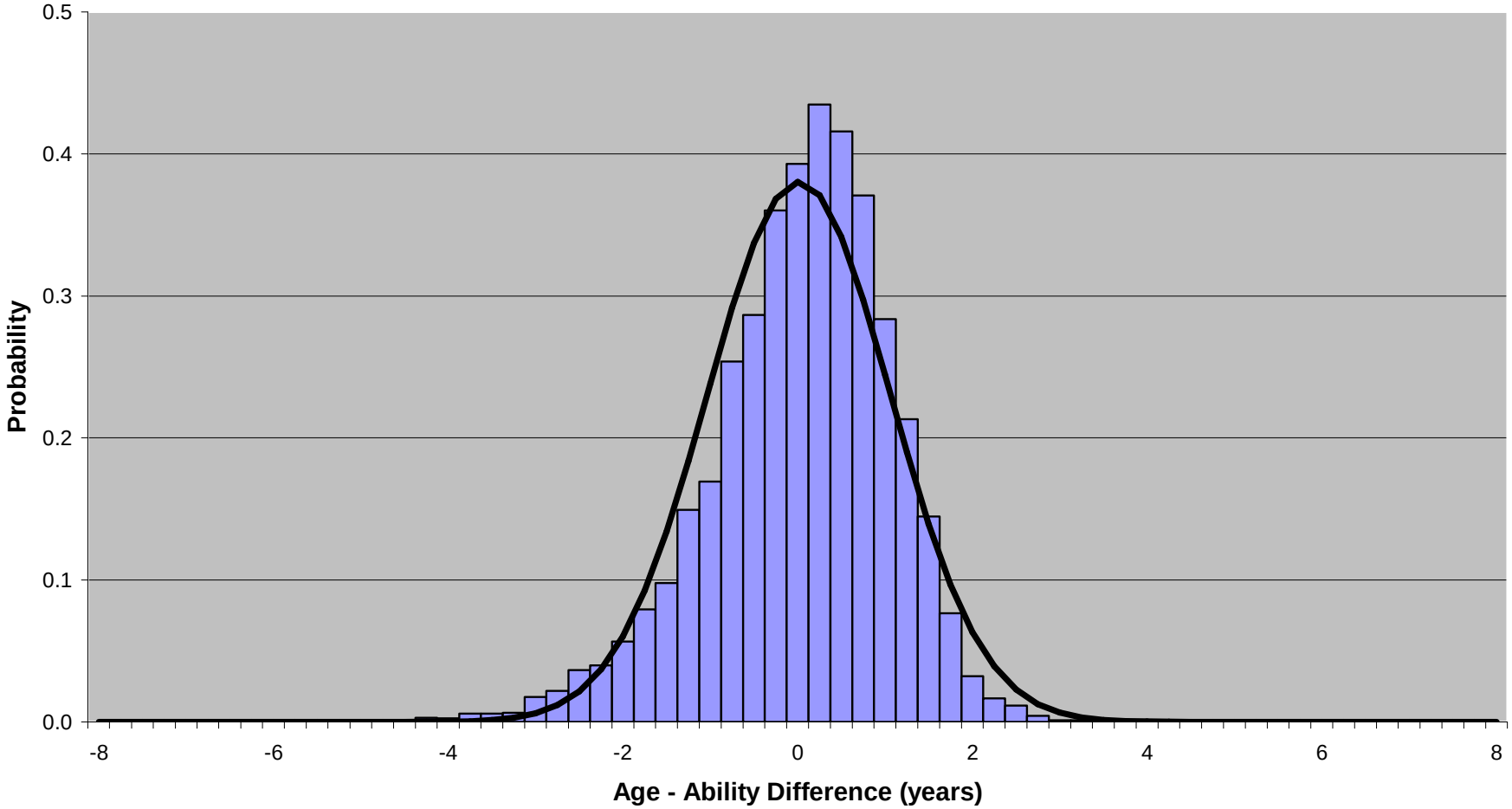


Figure 2.3.3.1: Normal Model Plot for Boys' Mathematics Results in P6

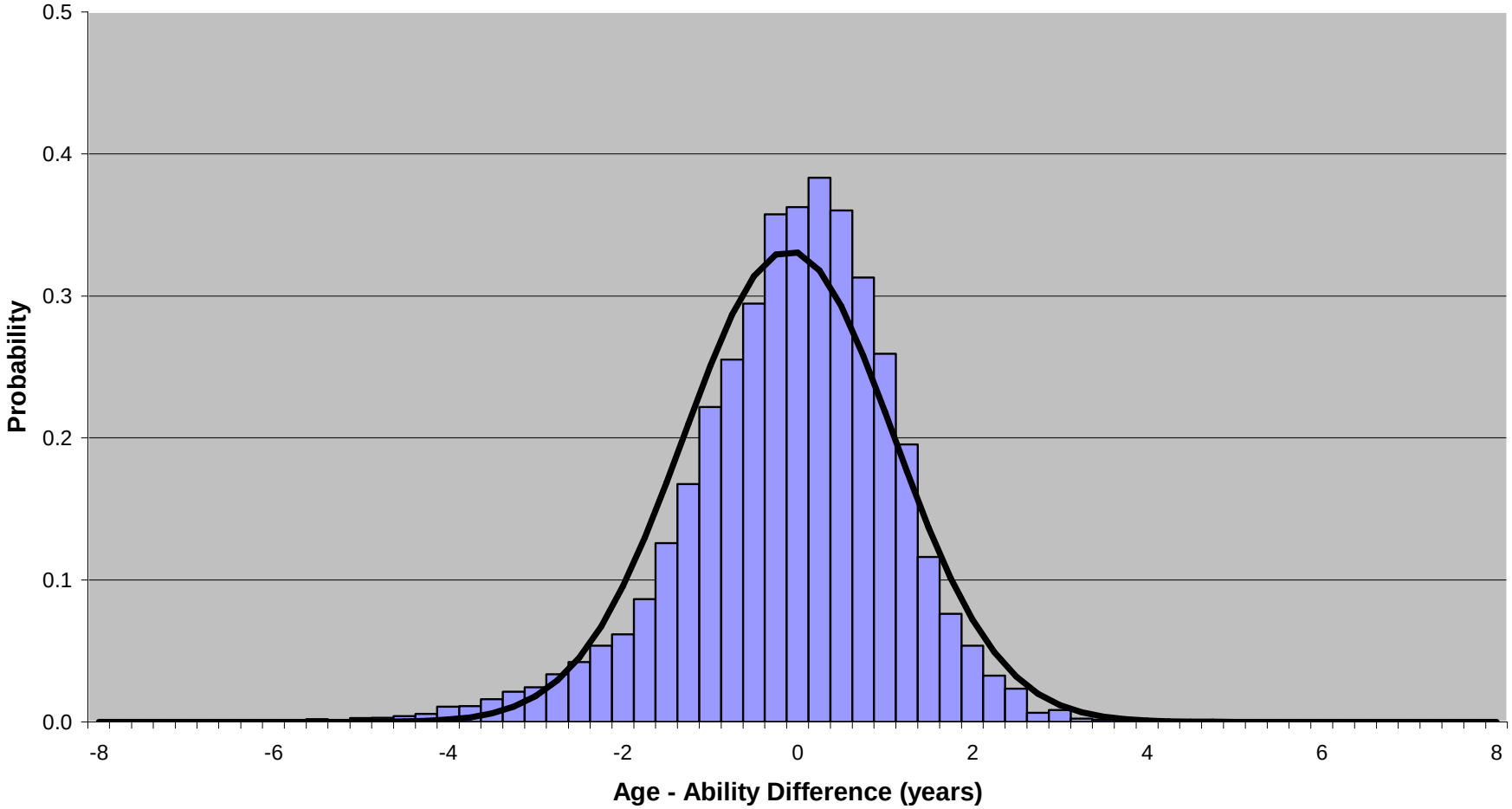


Figure 2.3.3.2: Normal Model Plot for Girls' Mathematics Results in P6

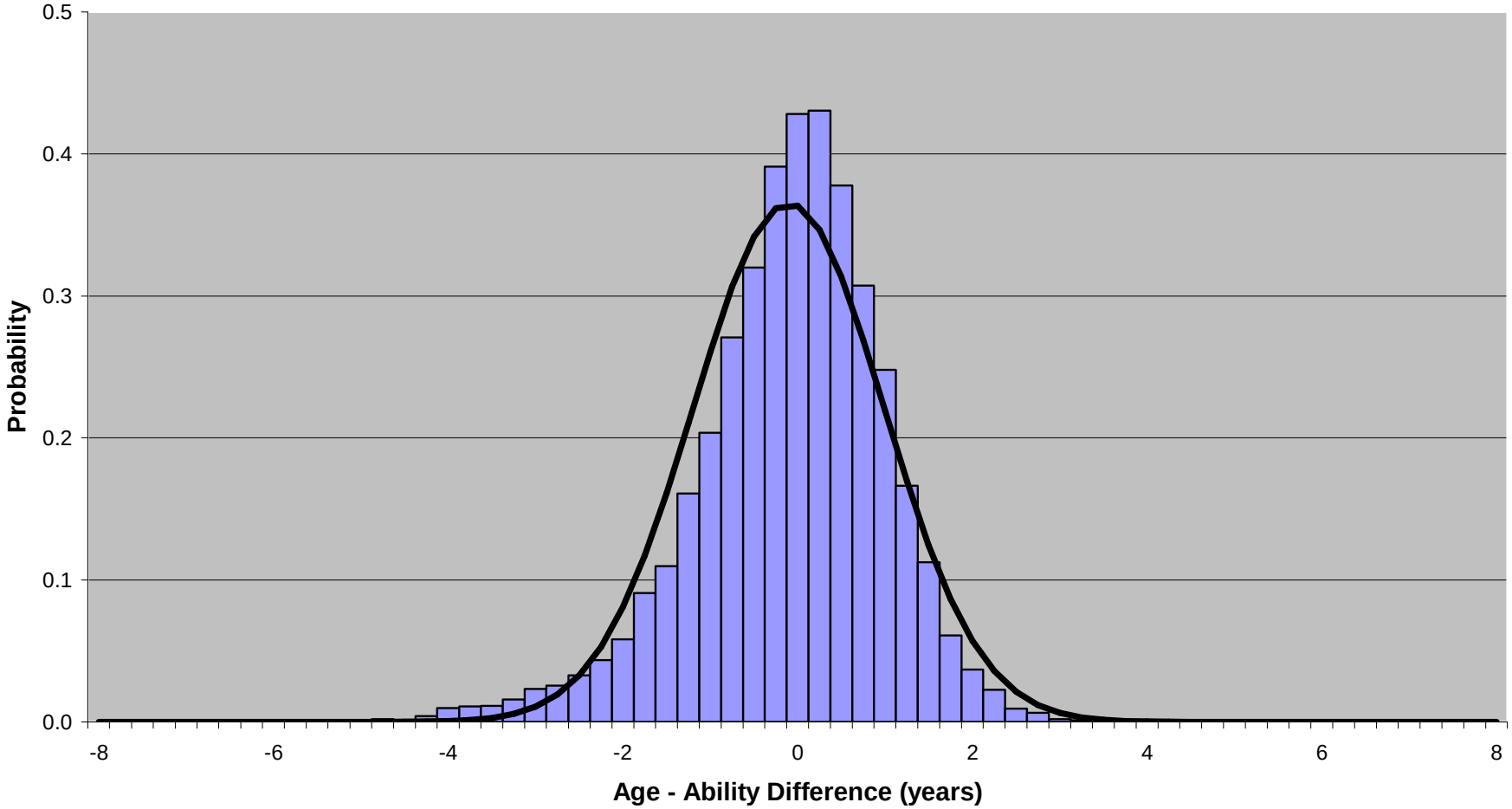


Figure 2.3.4.1: Normal Model Plot for Boys' Mathematics Results in P7

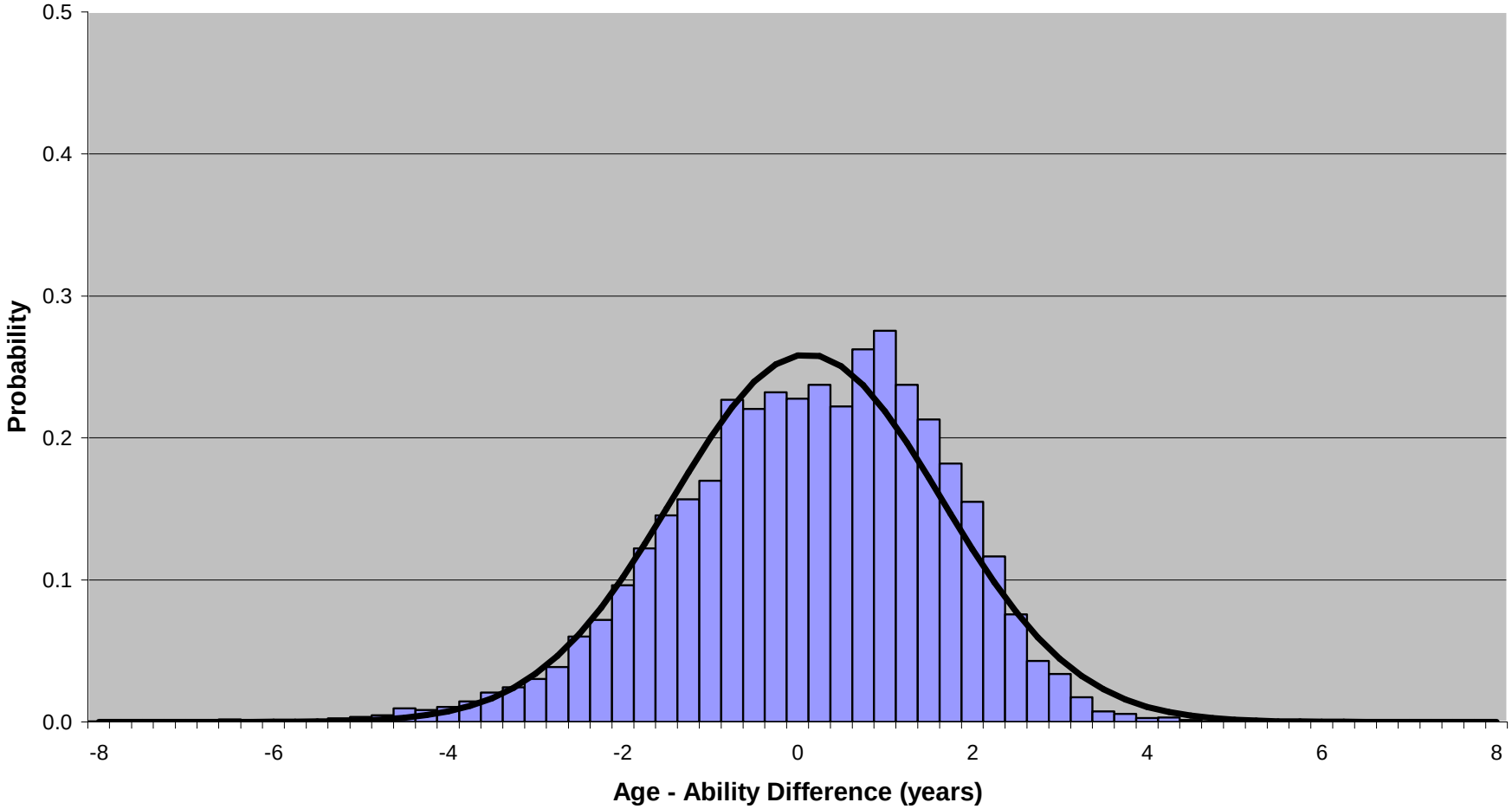


Figure 2.3.4.2: Normal Model Plot for Girls' Mathematics Results in P7

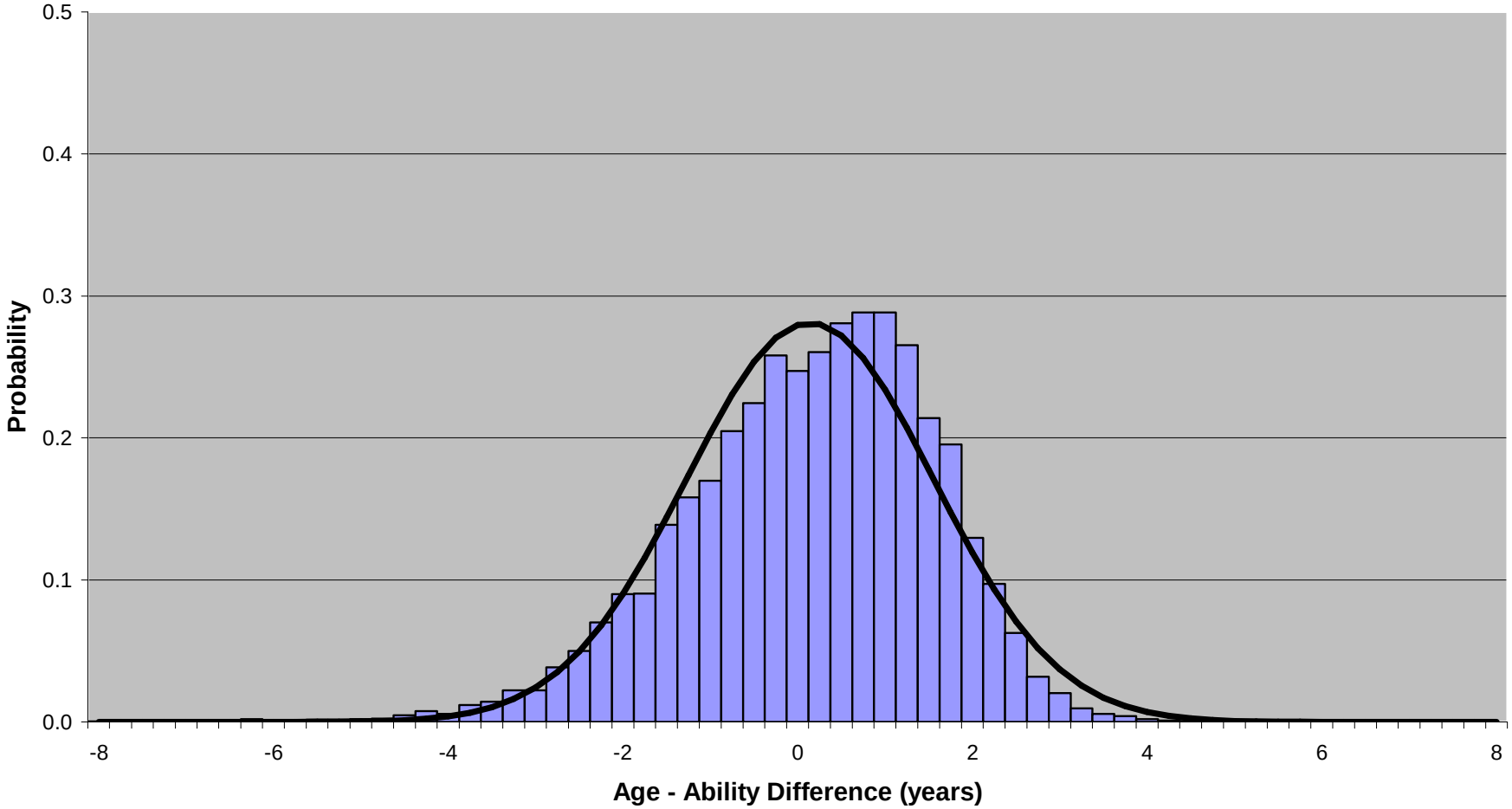


Figure 2.4.1.1: Normal Model Plot for Boys' Arithmetic Results in P4

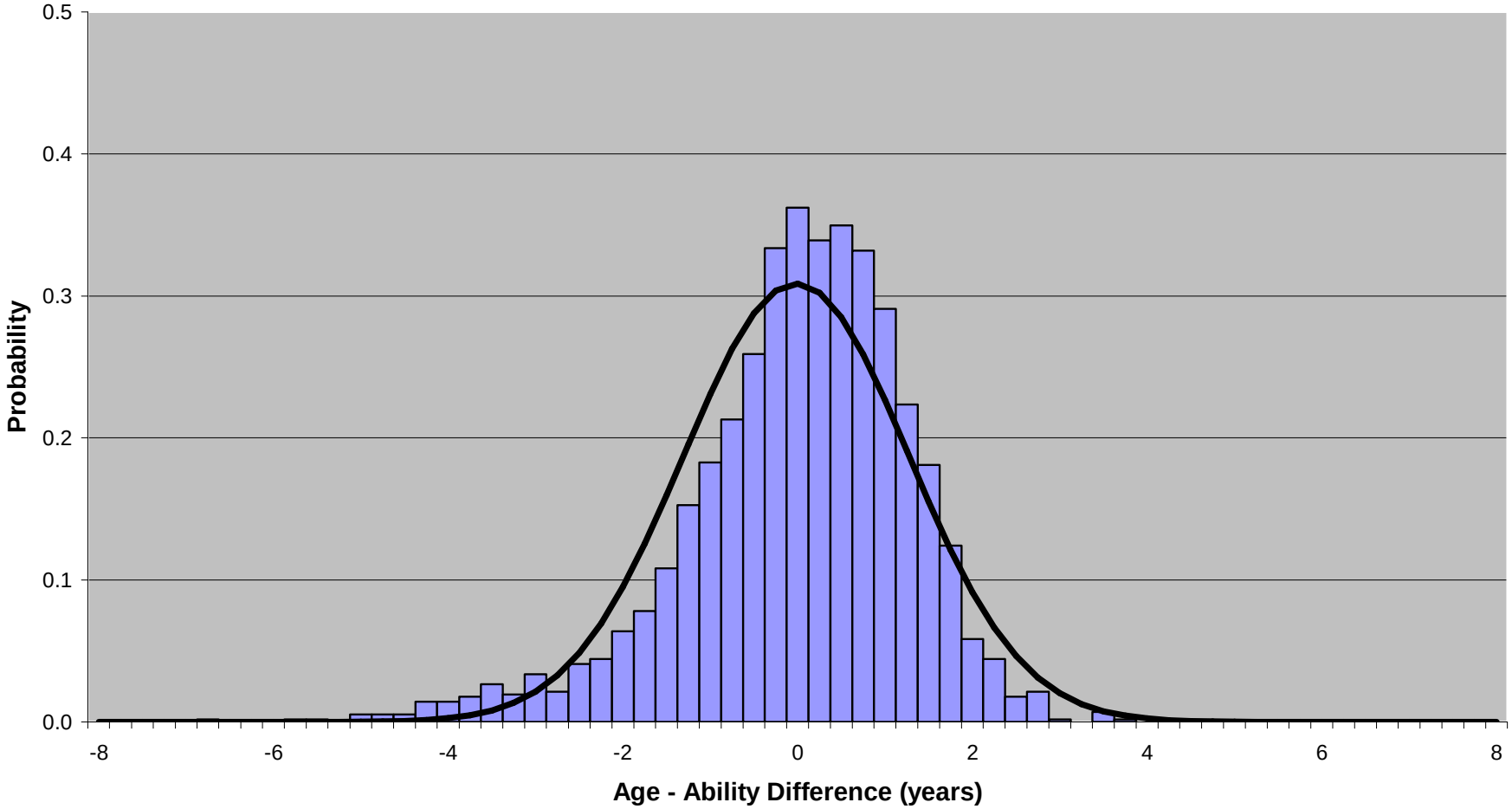


Figure 2.4.1.2: Normal Model Plot for Girls' Arithmetic Results in P4

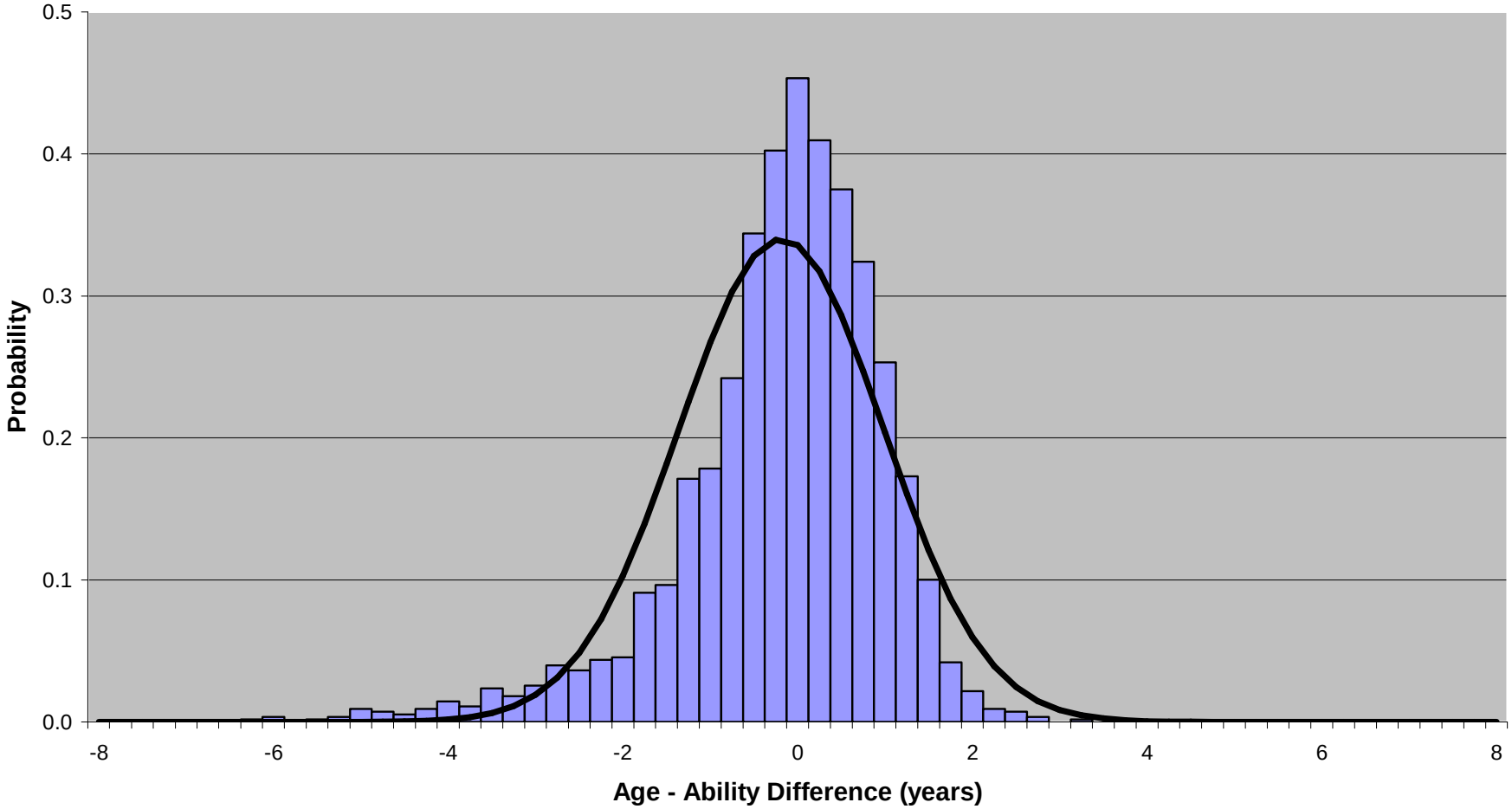


Figure 2.4.2.1: Normal Model Plot for Boys' Arithmetic Results in P5

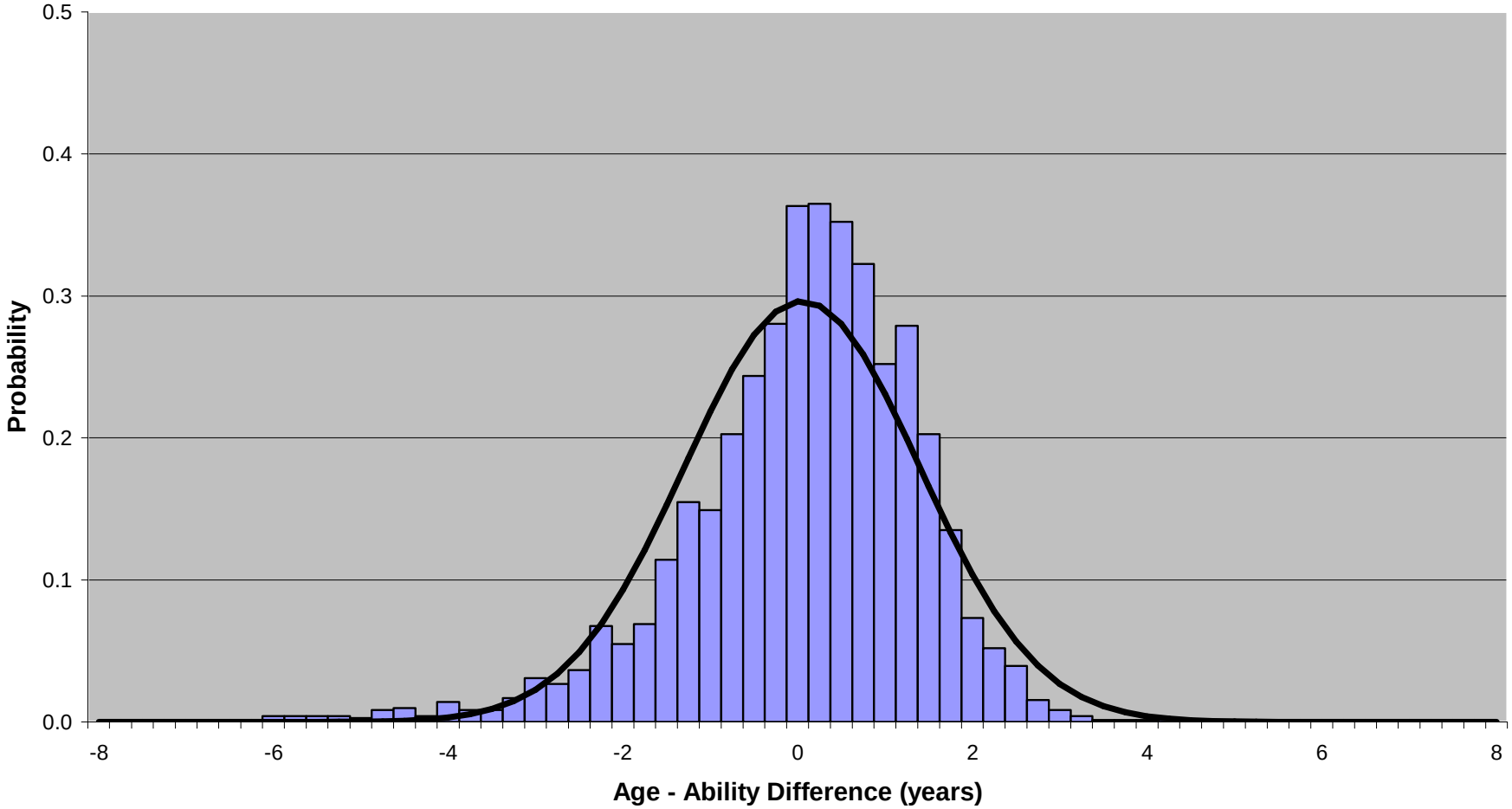


Figure 2.4.2.2: Normal Model Plot for Girls' Arithmetic Results in P5

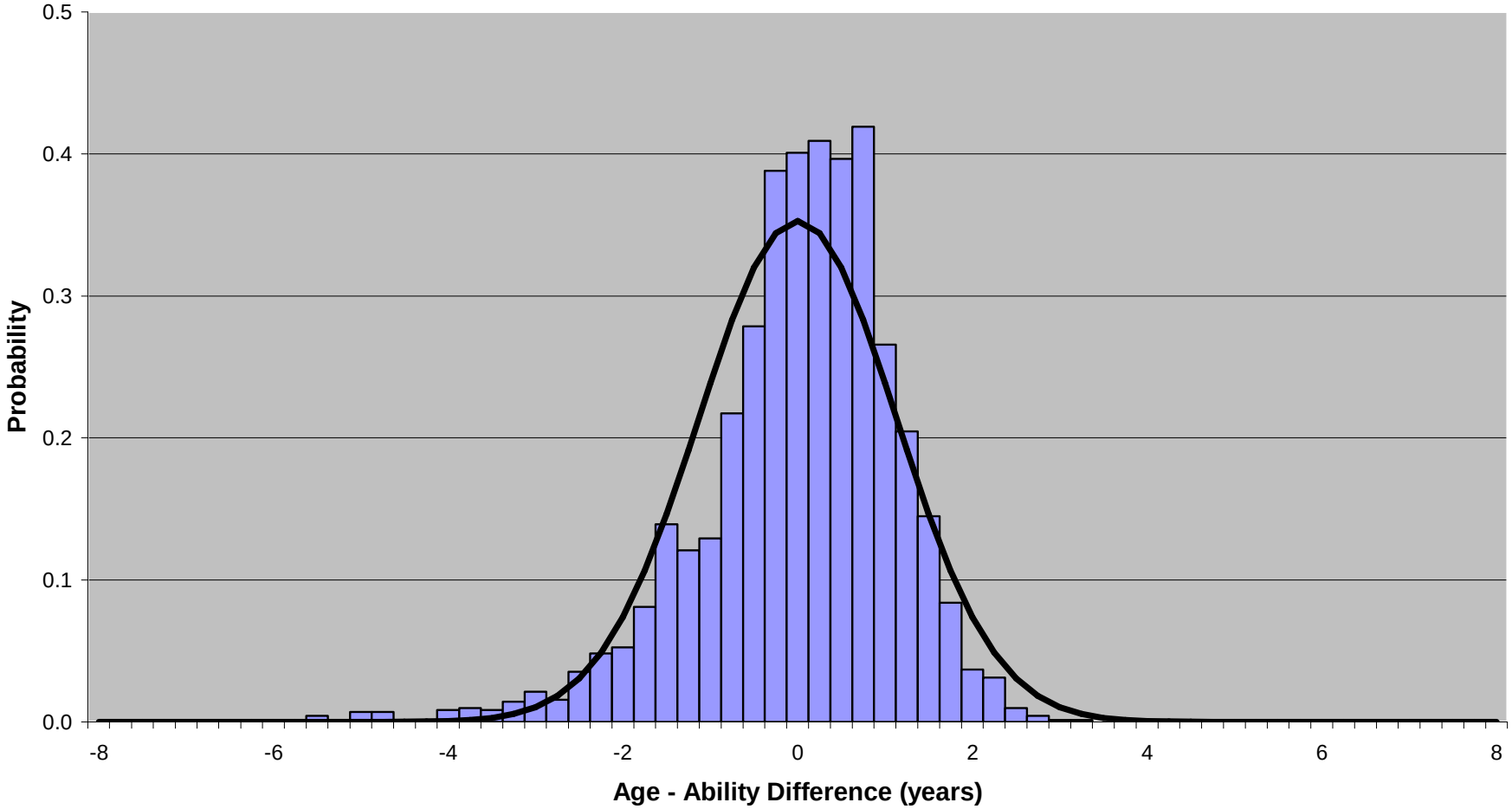


Figure 2.4.3.1: Normal Model Plot for Boys' Arithmetic Results in P6

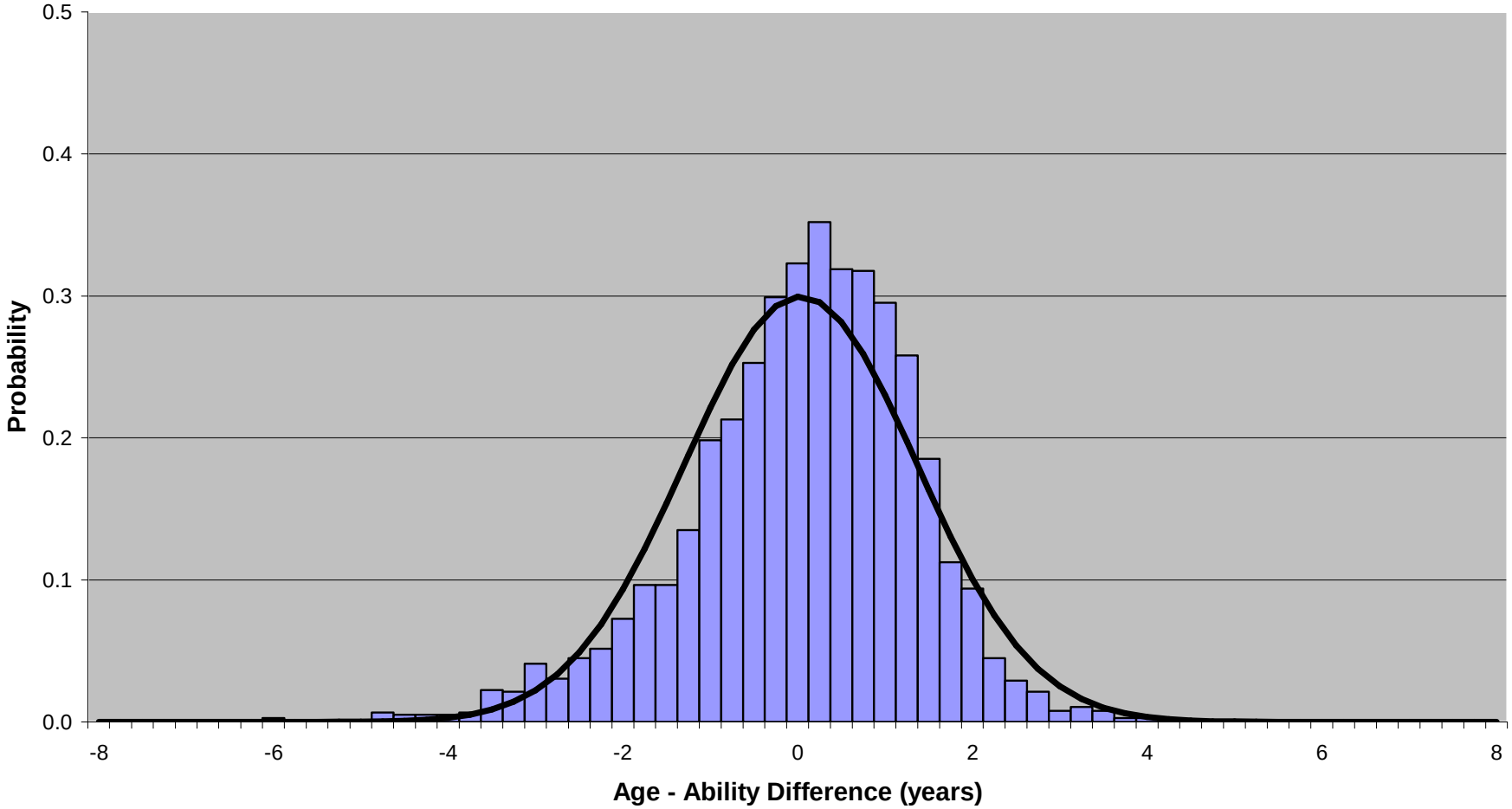


Figure 2.4.3.2: Normal Model Plot for Girls' Arithmetic Results in P6

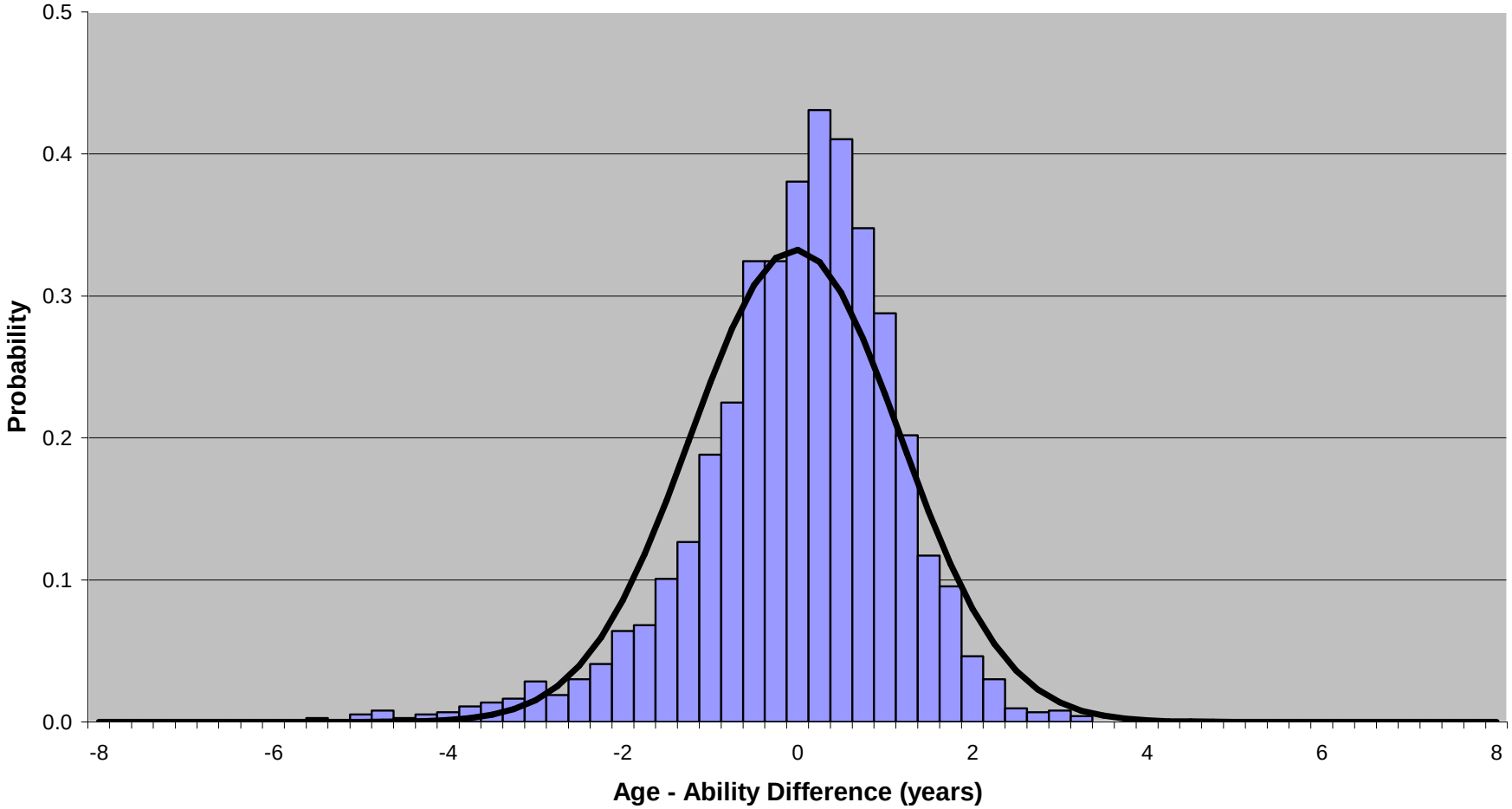


Figure 2.4.4.1: Normal Model Plot for Boys' Arithmetic Results in P7

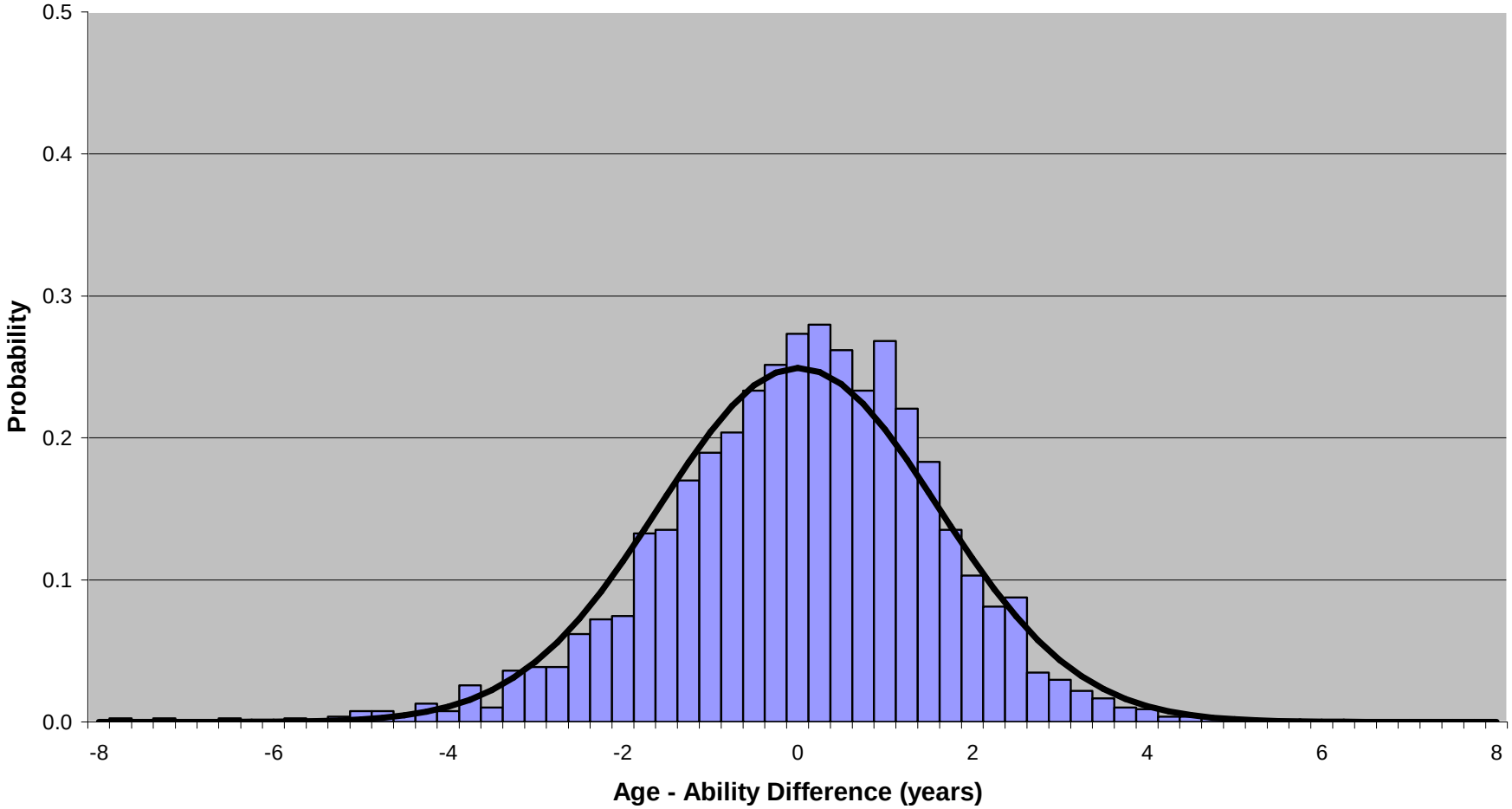


Figure 2.4.4.2: Normal Model Plot for Girls' Arithmetic Results in P7

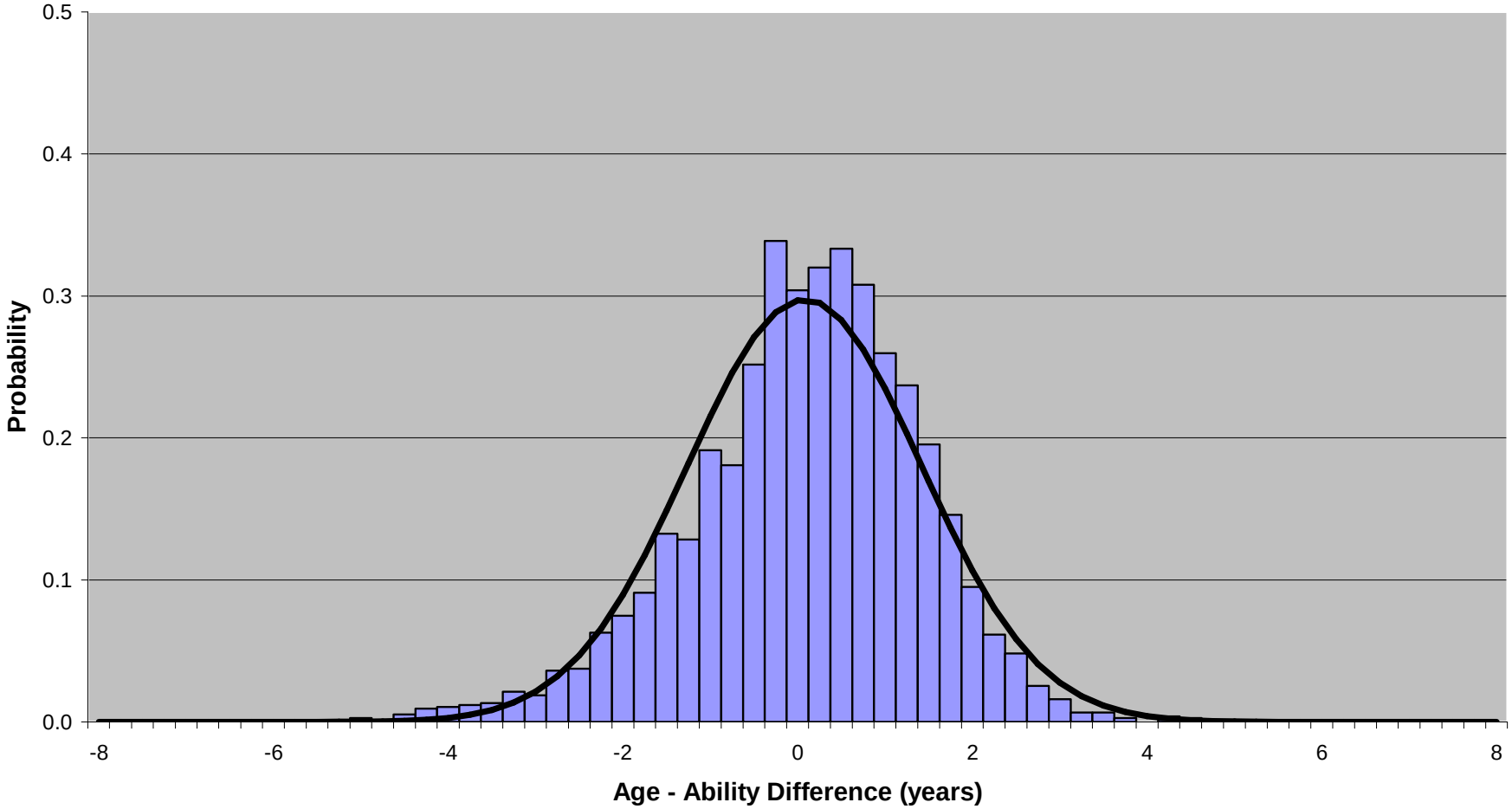


Figure 3.1.1.1: Binormal Model Plot for Boys' Picture Vocabulary Results in P4

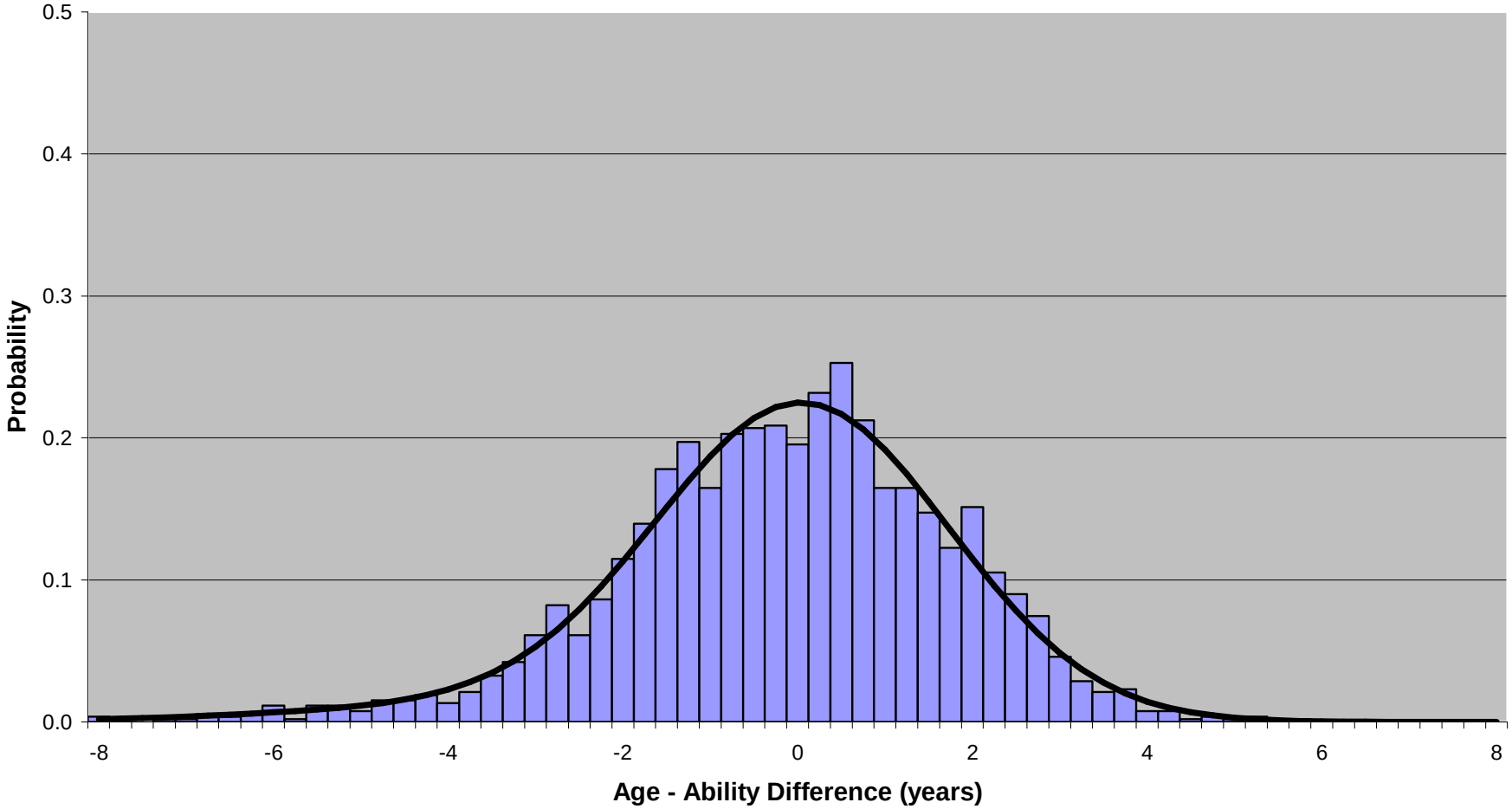


Figure 3.1.1.2: Binormal Model Plot for Girls' Picture Vocabulary Results in P4

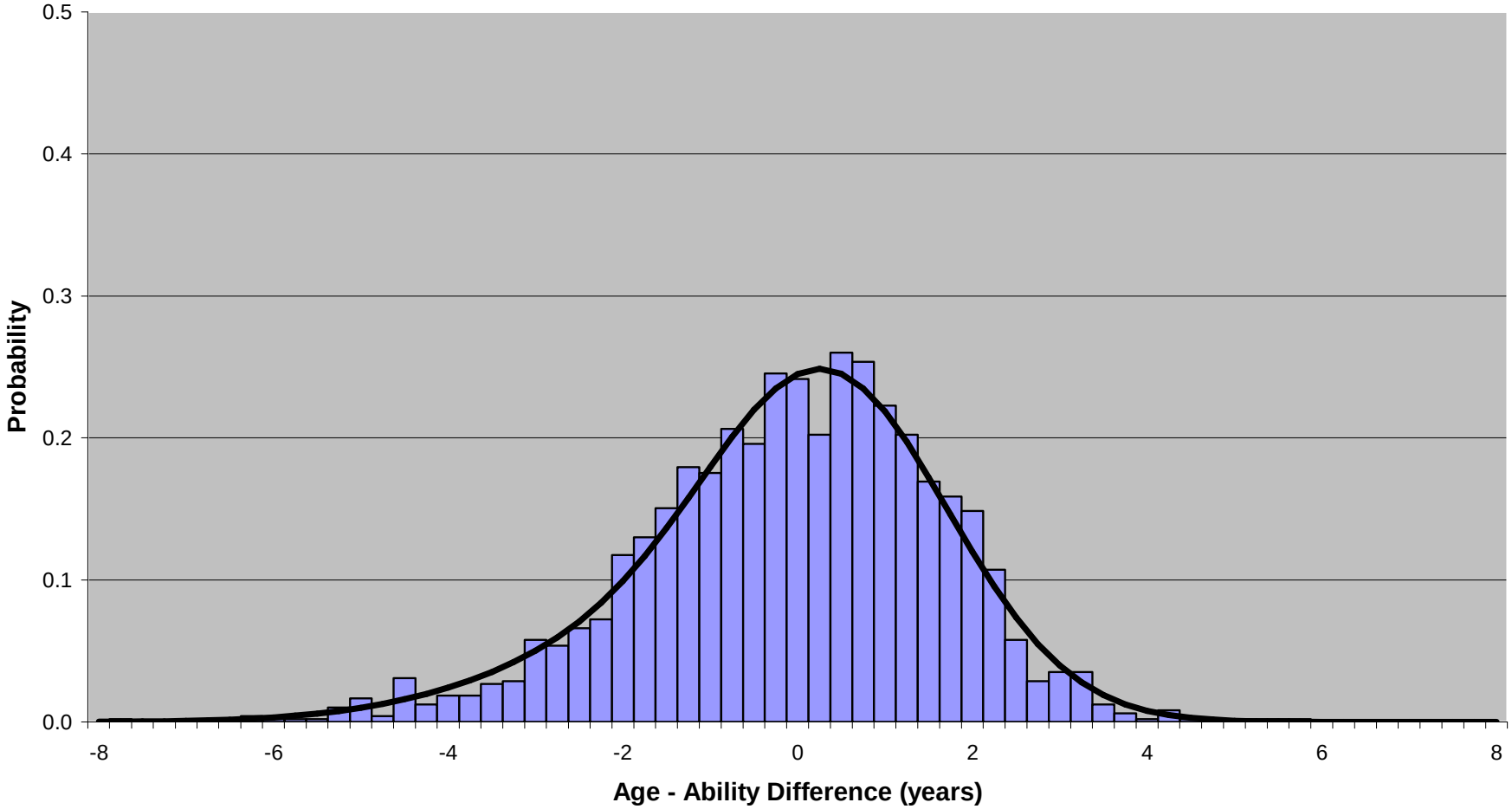


Figure 3.1.2.1: Binormal Model Plot for Boys' Picture Vocabulary Results in P5

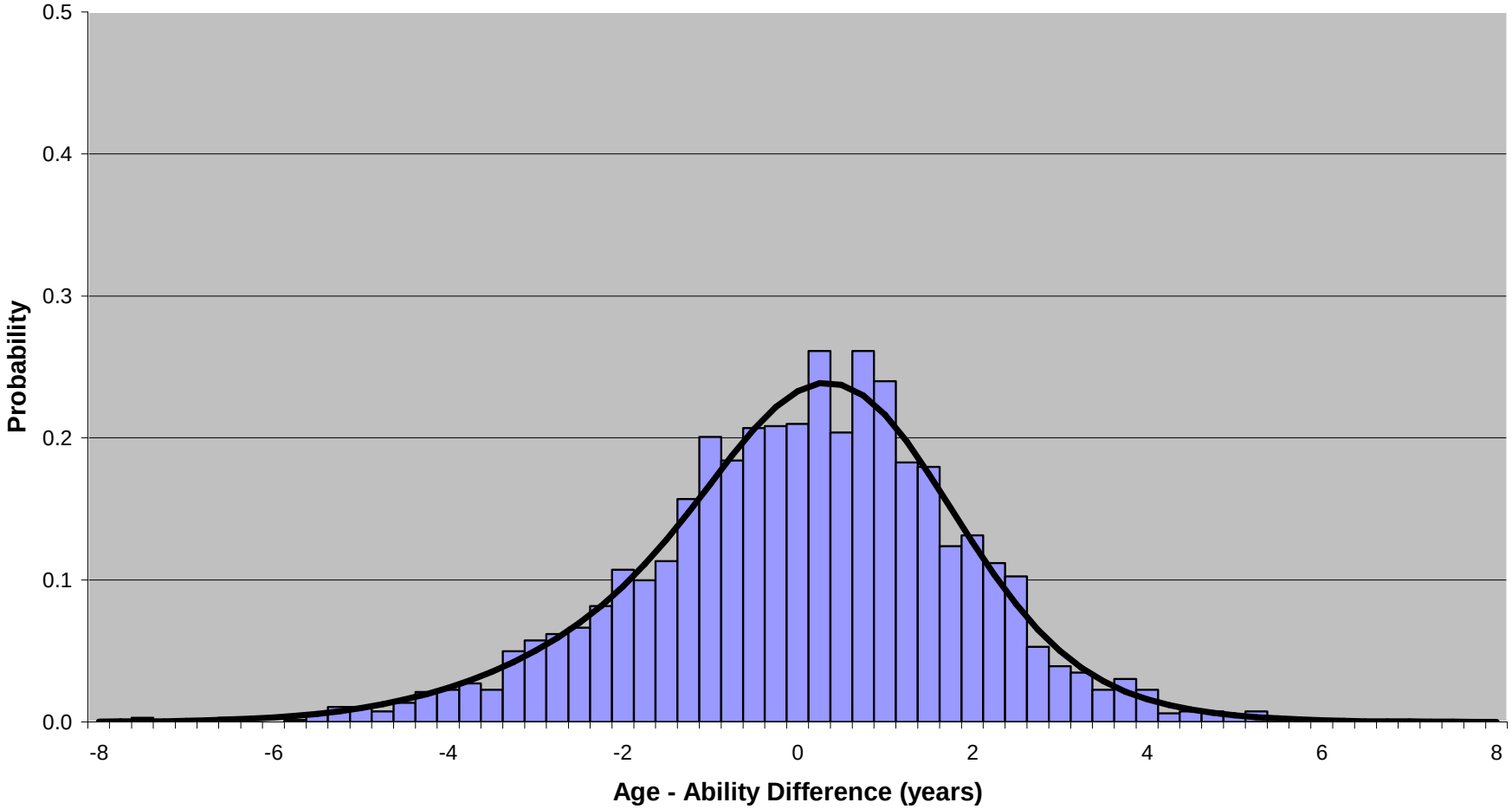


Figure 3.1.2.2: Binormal Model Plot for Girls' Picture Vocabulary Results in P5

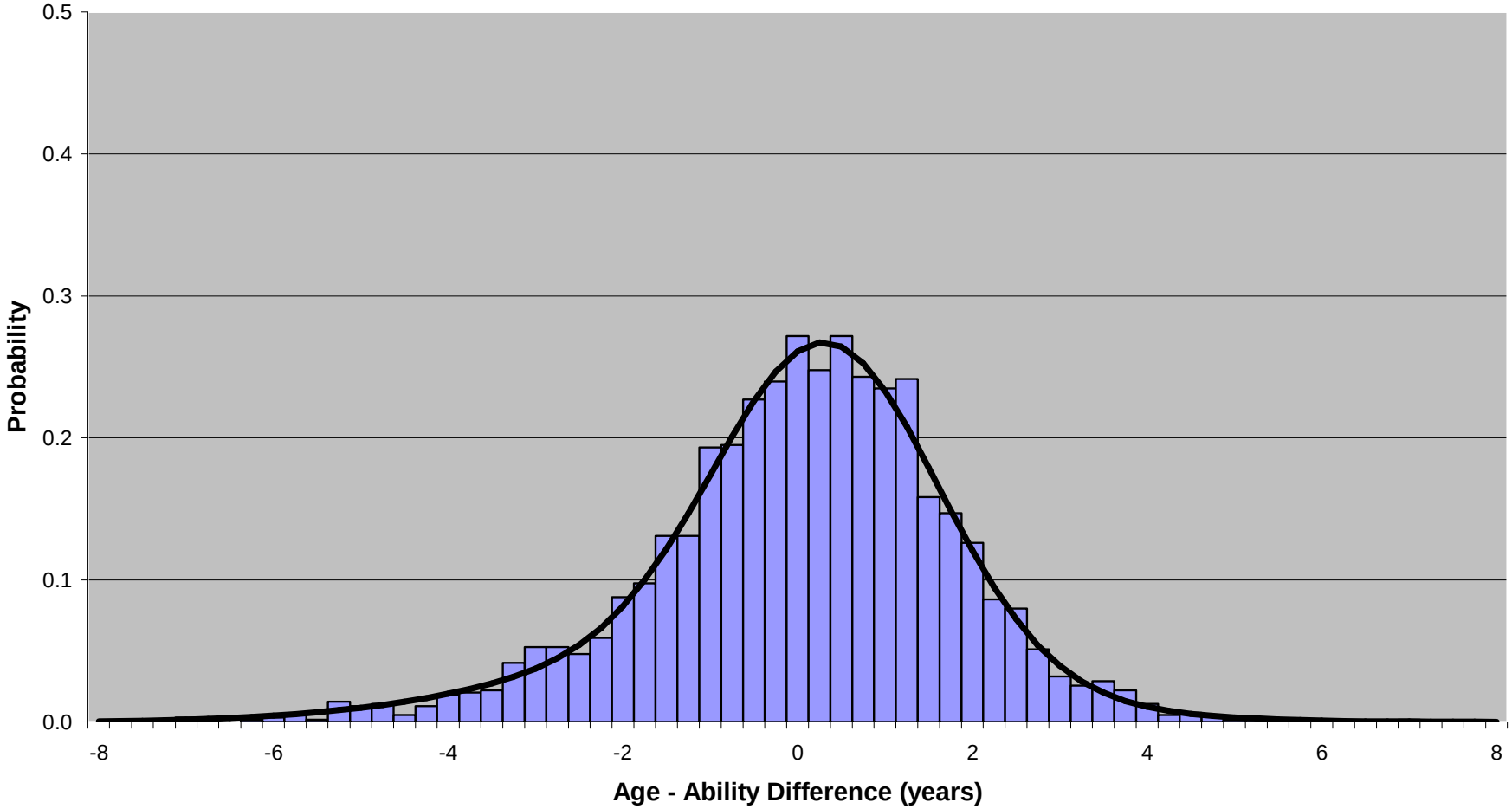


Figure 3.1.3.1: Binormal Model Plot for Boys' Picture Vocabulary Results in P6

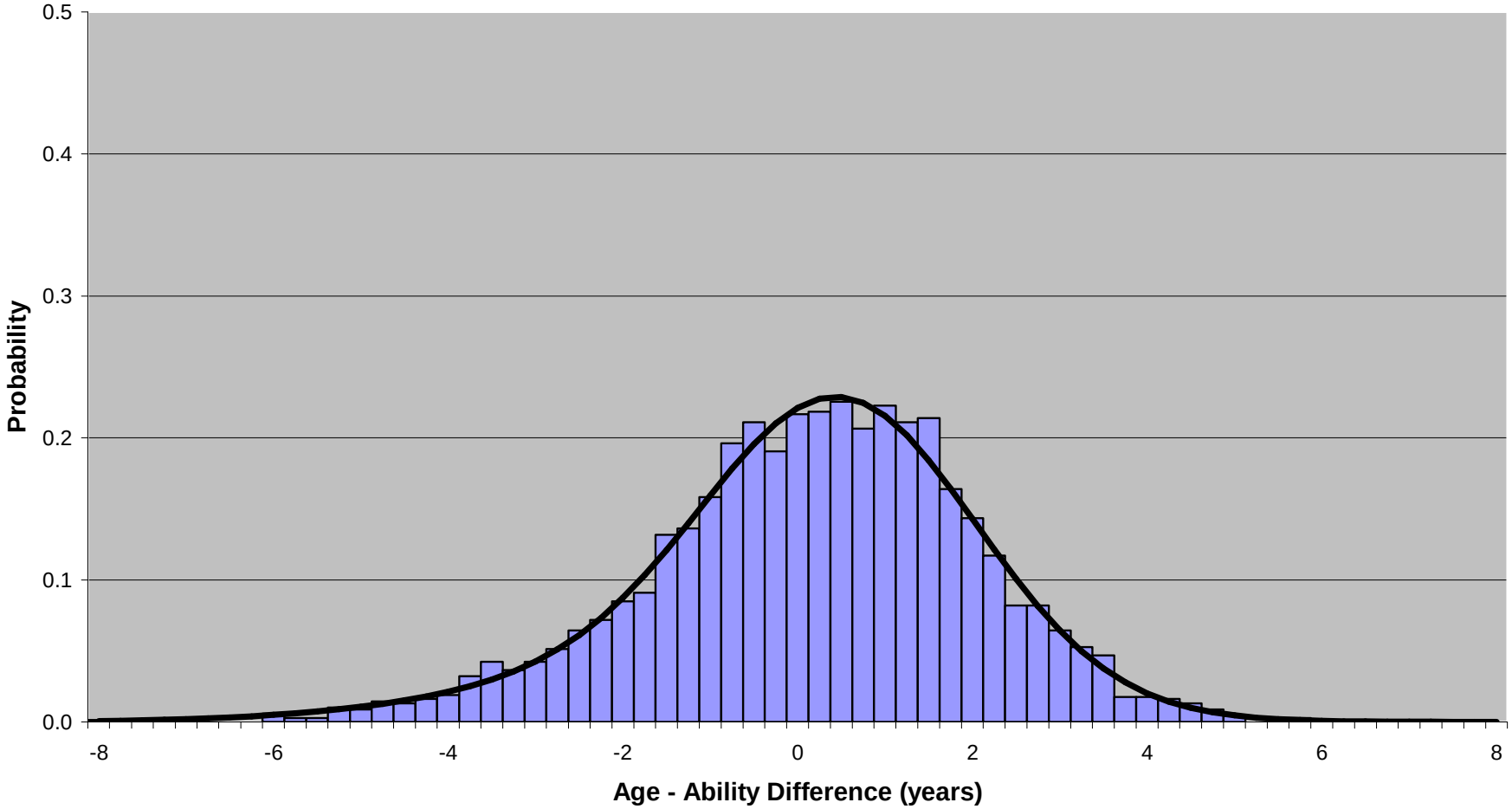


Figure 3.1.3.2: Binormal Model Plot for Girls' Picture Vocabulary Results in P6

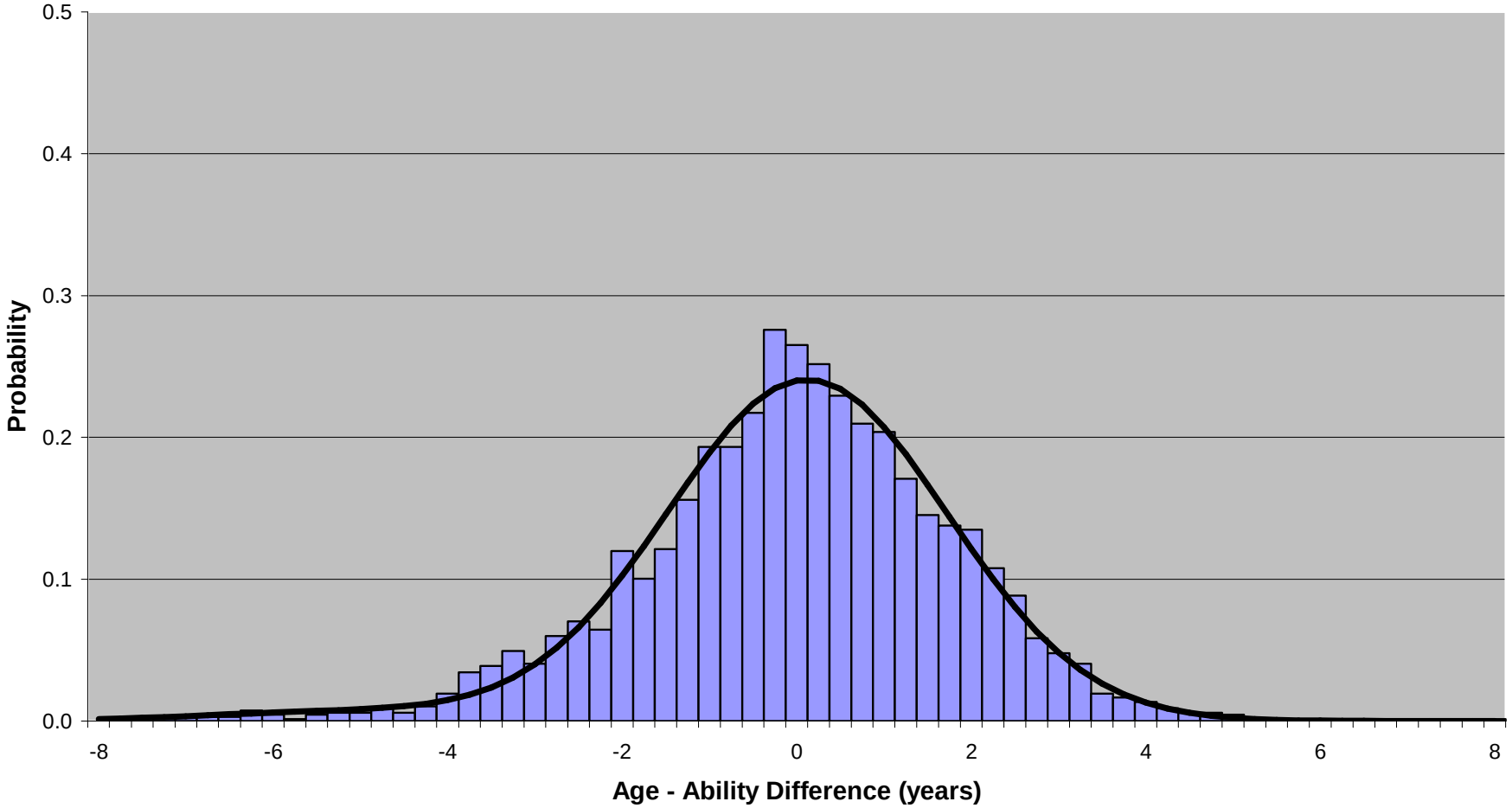


Figure 3.1.4.1: Binormal Model Plot for Boys' Picture Vocabulary Results in P7

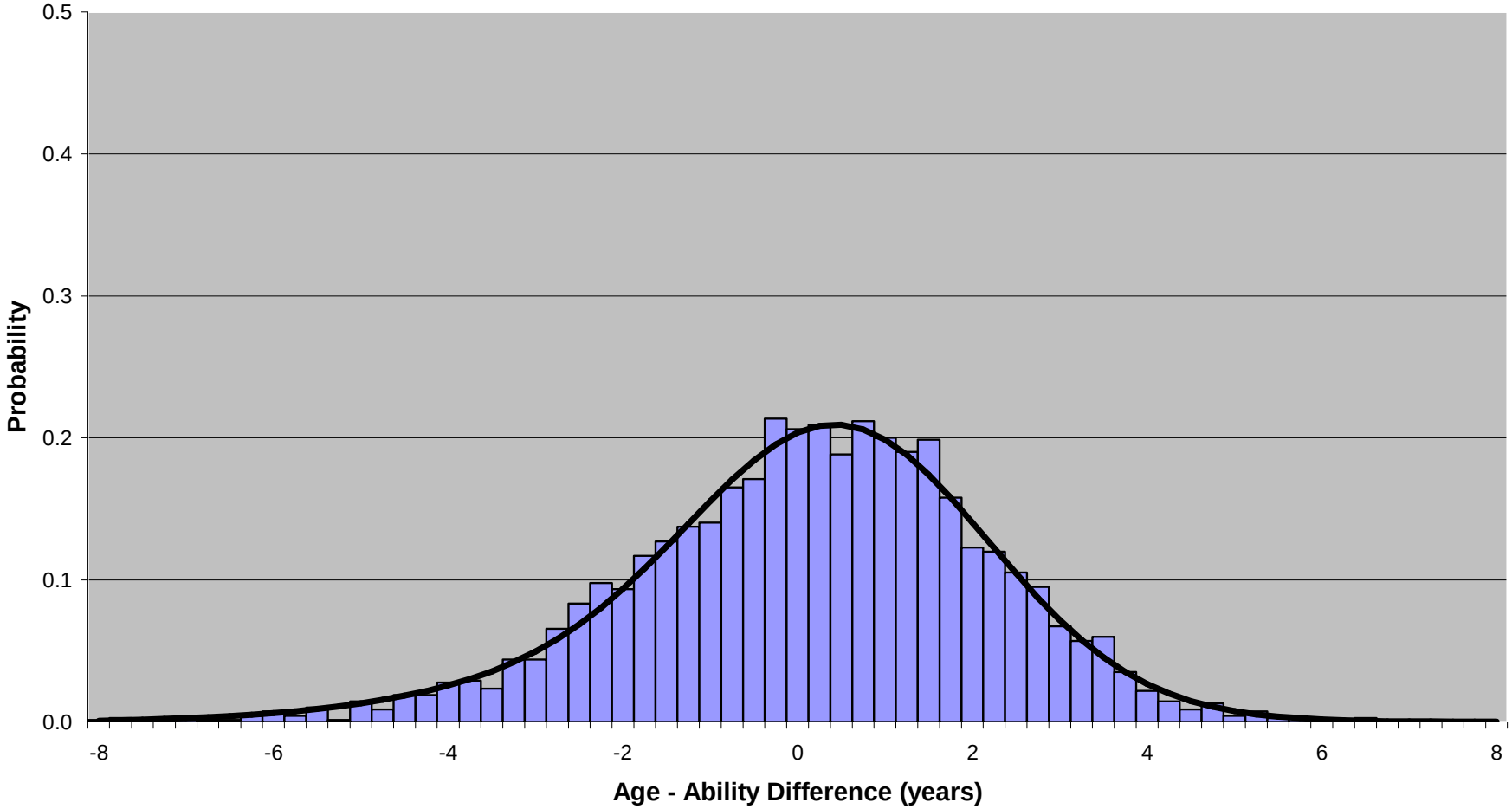


Figure 3.1.4.2: Binormal Model Plot for Girls' Picture Vocabulary Results in P7

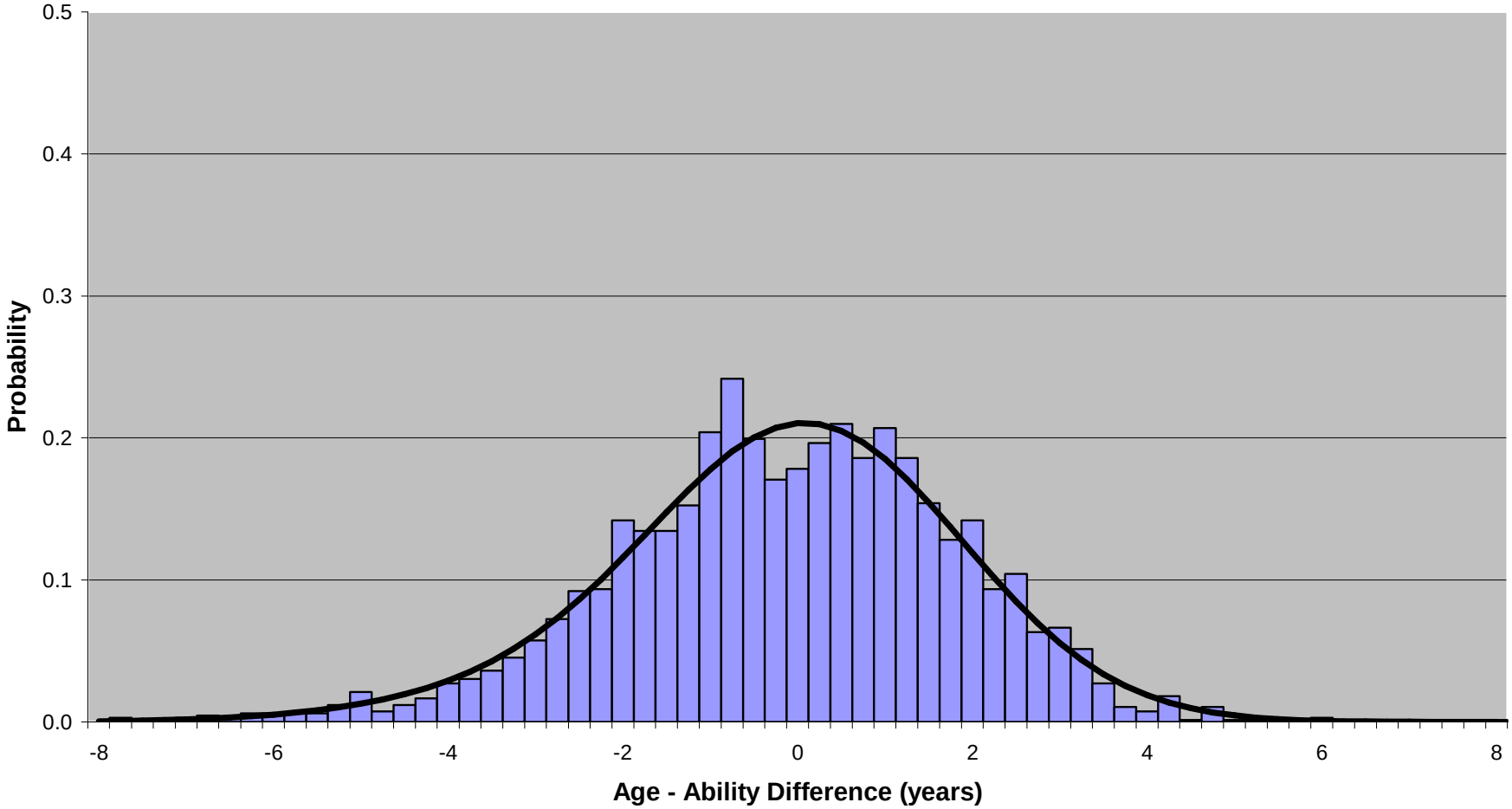


Figure 3.2.1.1: Binormal Model Plot for Boys' Reading Results in P4

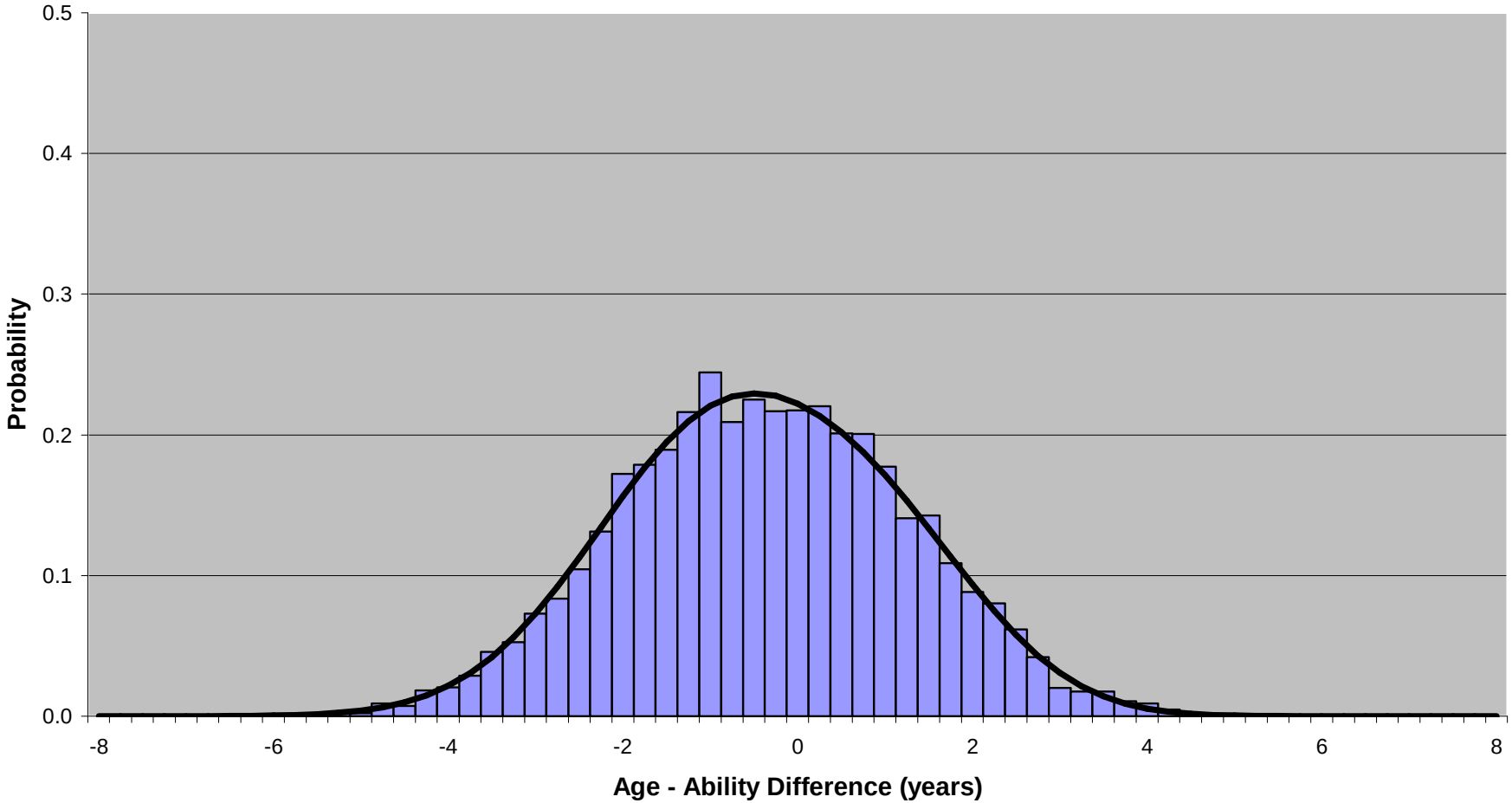


Figure 3.2.1.2: Binormal Model Plot for Girls' Reading Results in P4

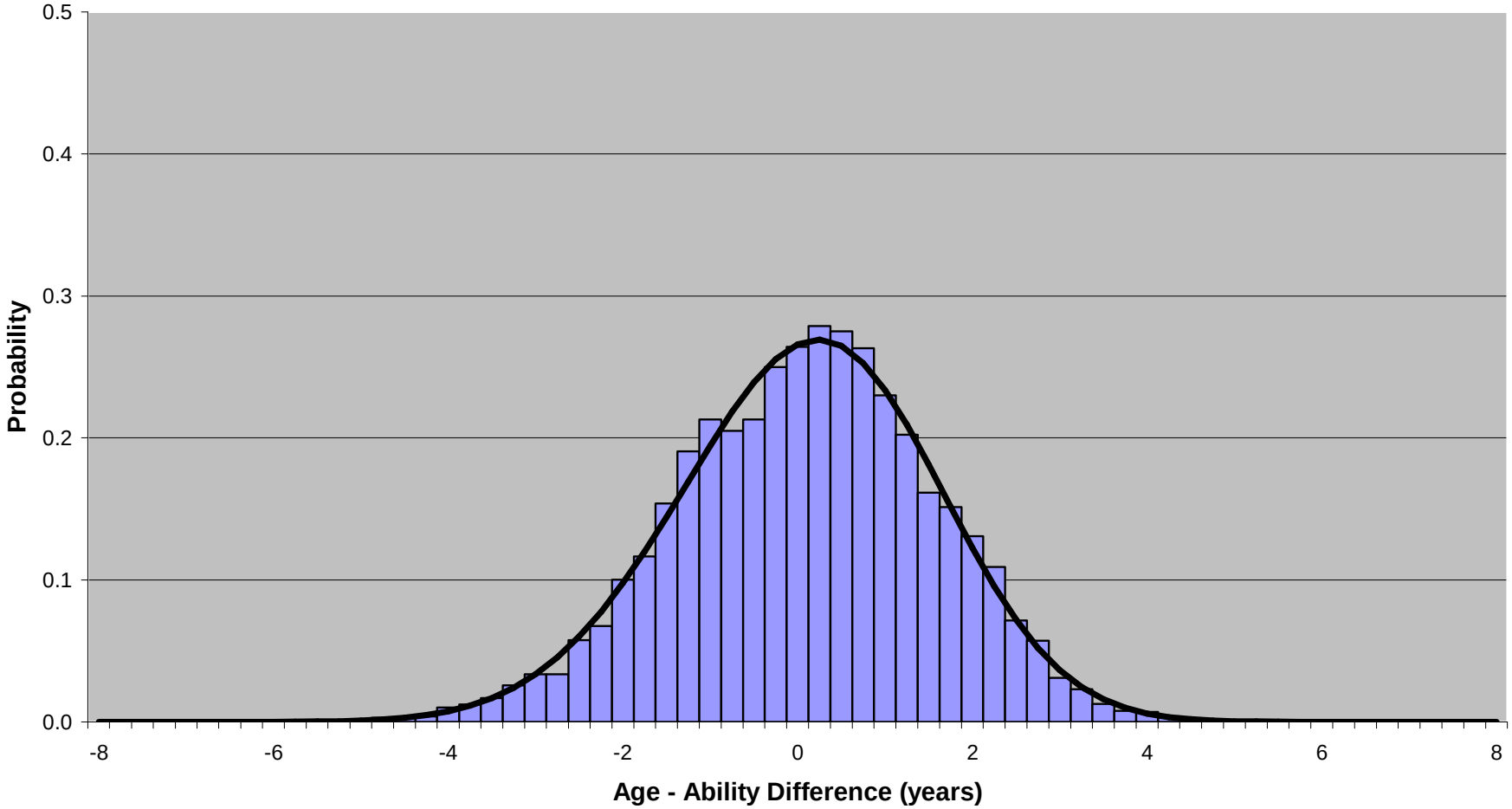


Figure 3.2.2.1: Binormal Model Plot for Boys' Reading Results in P5

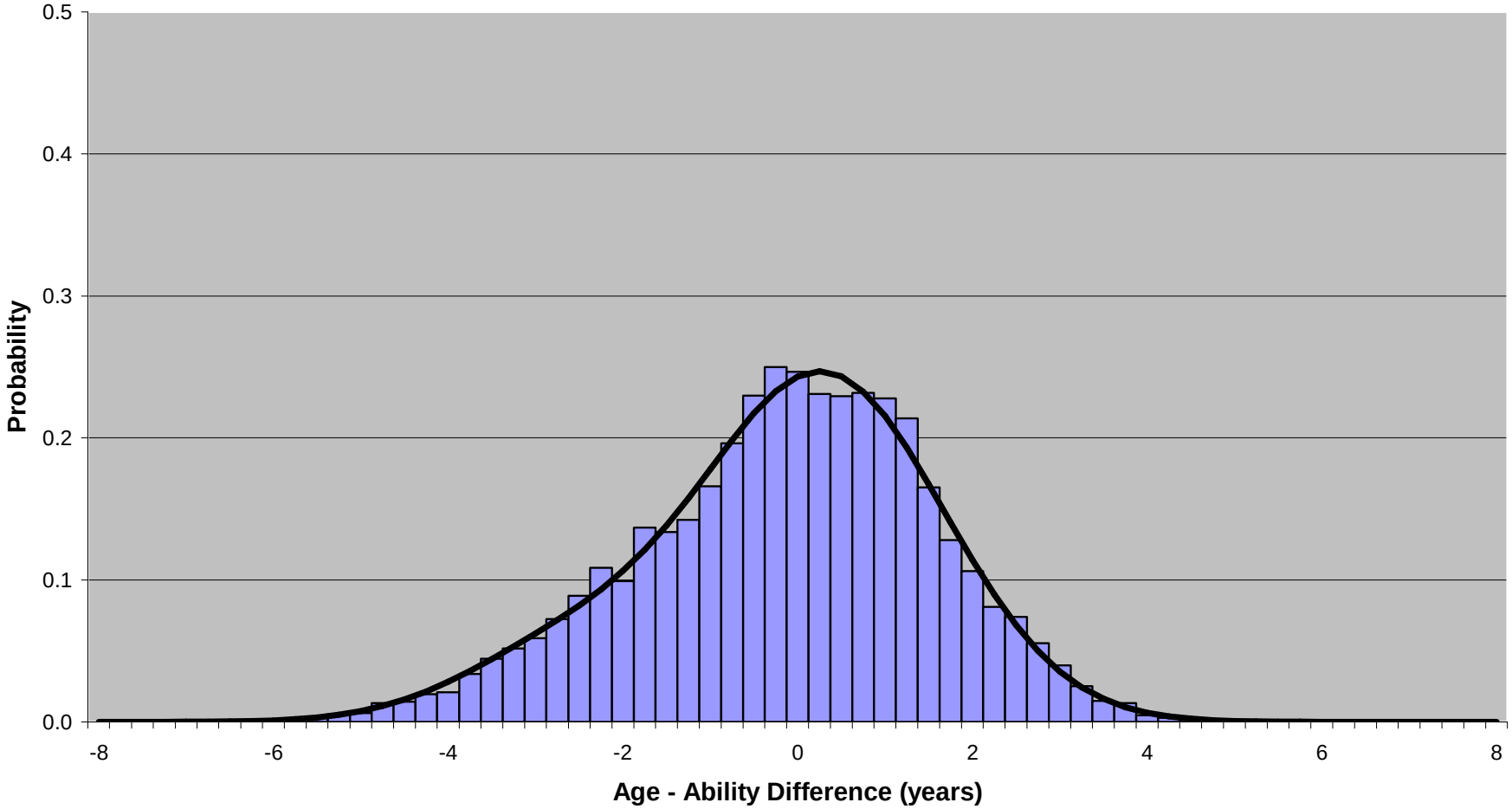


Figure 3.2.2.2: Binormal Model Plot for Girls' Reading Results in P5

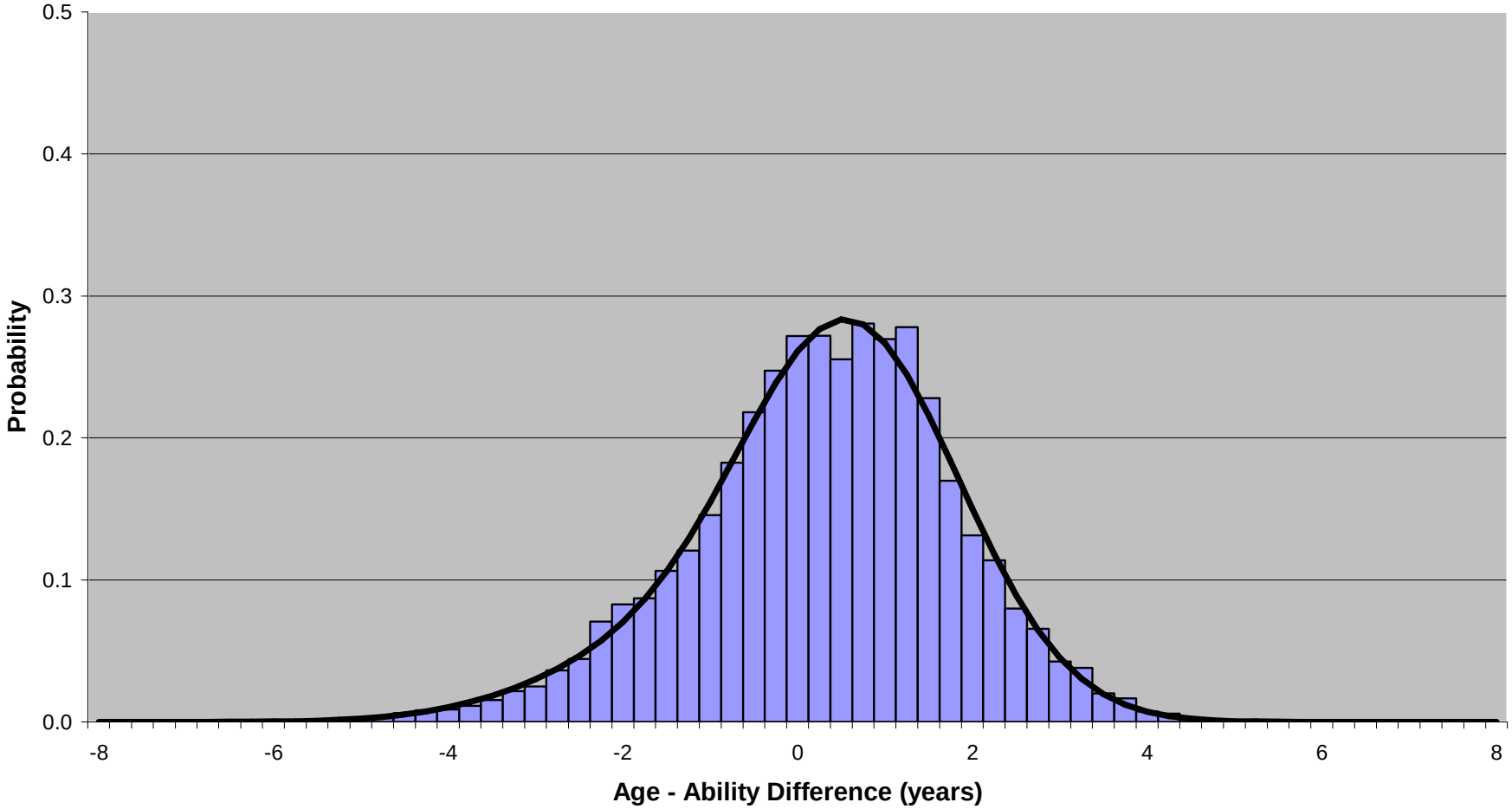


Figure 3.2.3.1: Binormal Model Plot for Boys' Reading Results in P6

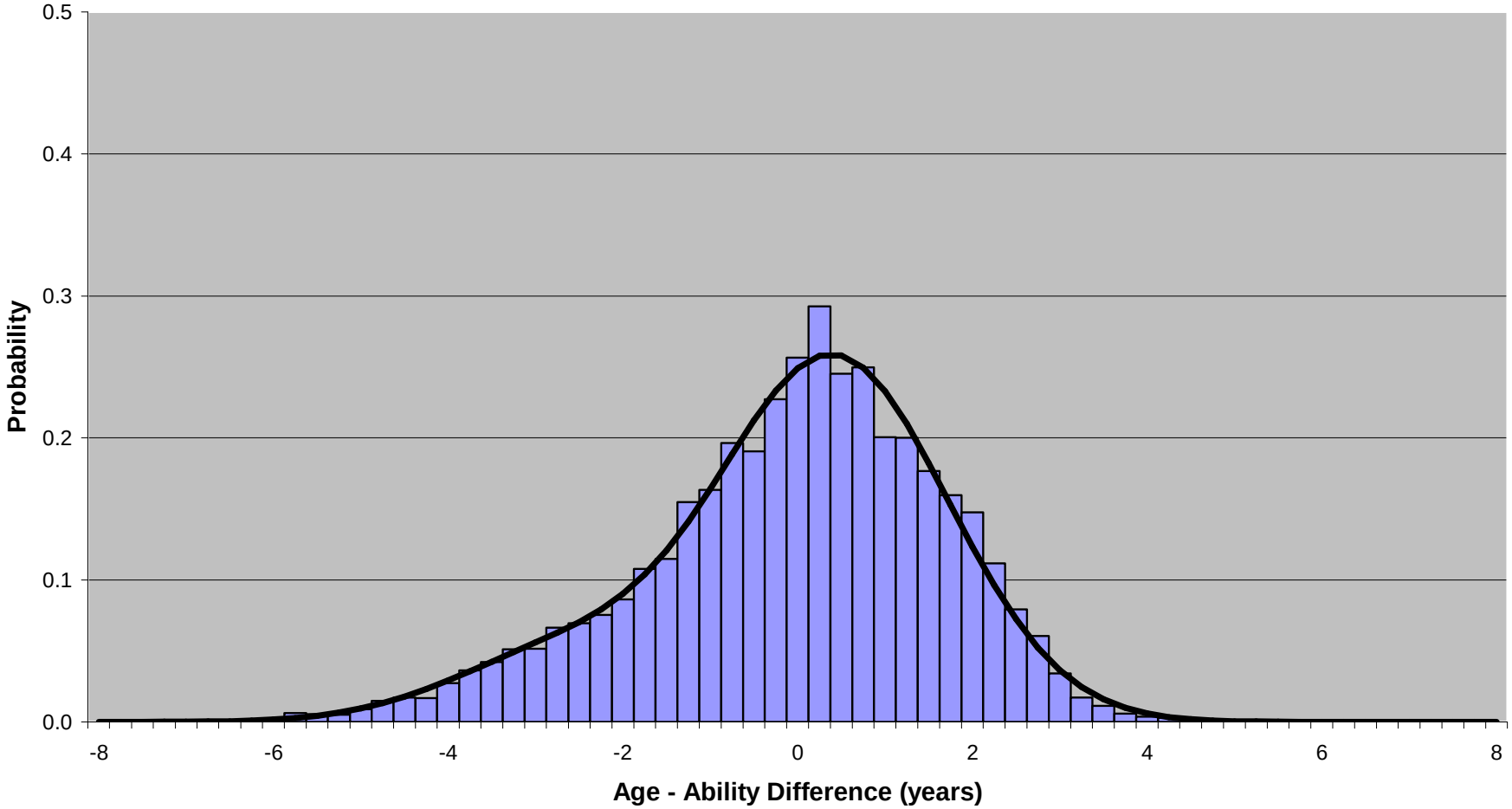


Figure 3.2.3.2: Binormal Model Plot for Girls' Reading Results in P6

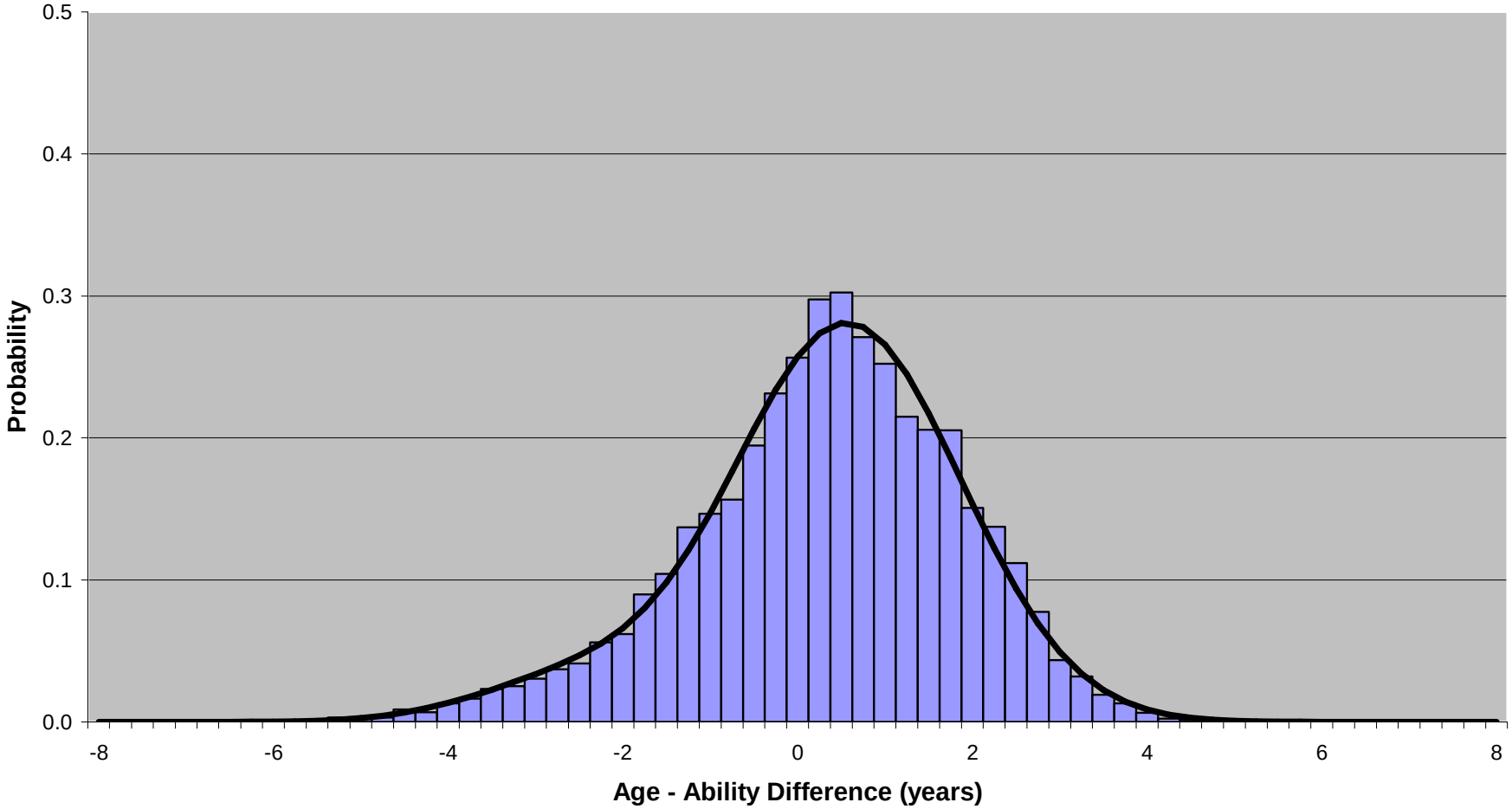


Figure 3.2.4.1: Binormal Model Plot for Boys' Reading Results in P7

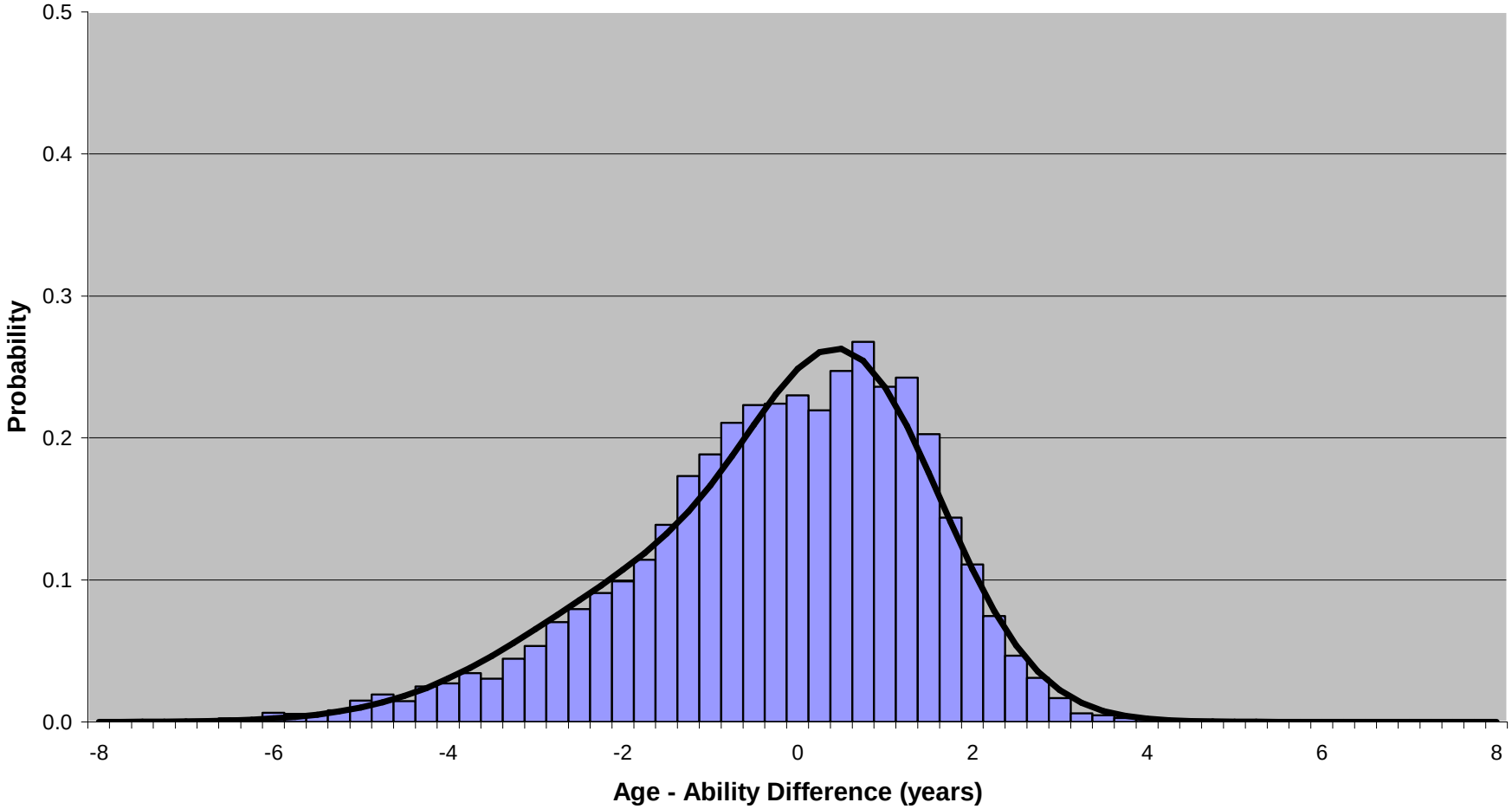


Figure 3.2.4.2: Binormal Model Plot for Girls' Reading Results in P7

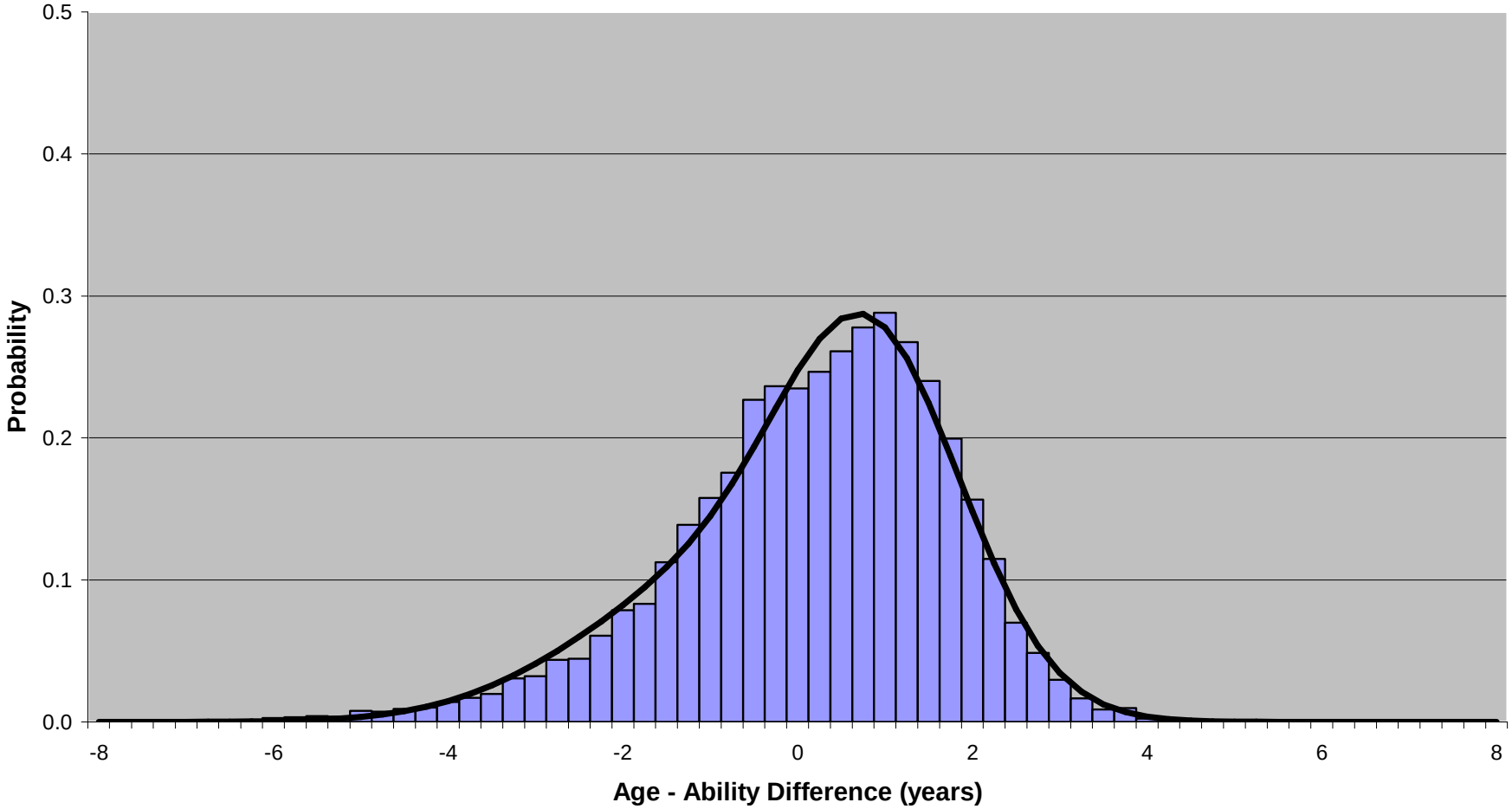


Figure 3.3.1.1: Binormal Model Plot for Boys' Mathematics Results in P4

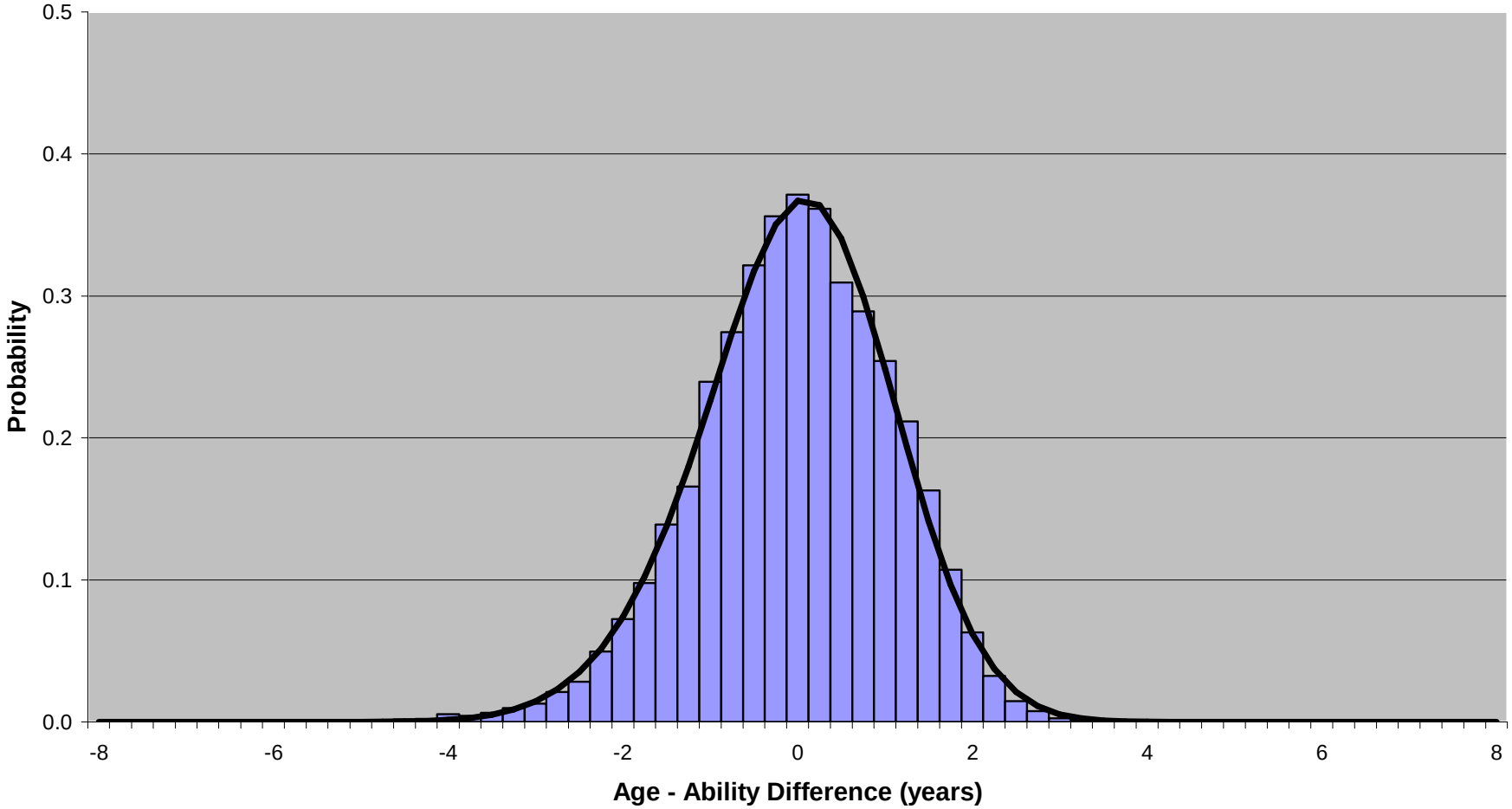


Figure 3.3.1.2: Binormal Model Plot for Girls' Mathematics Results in P4

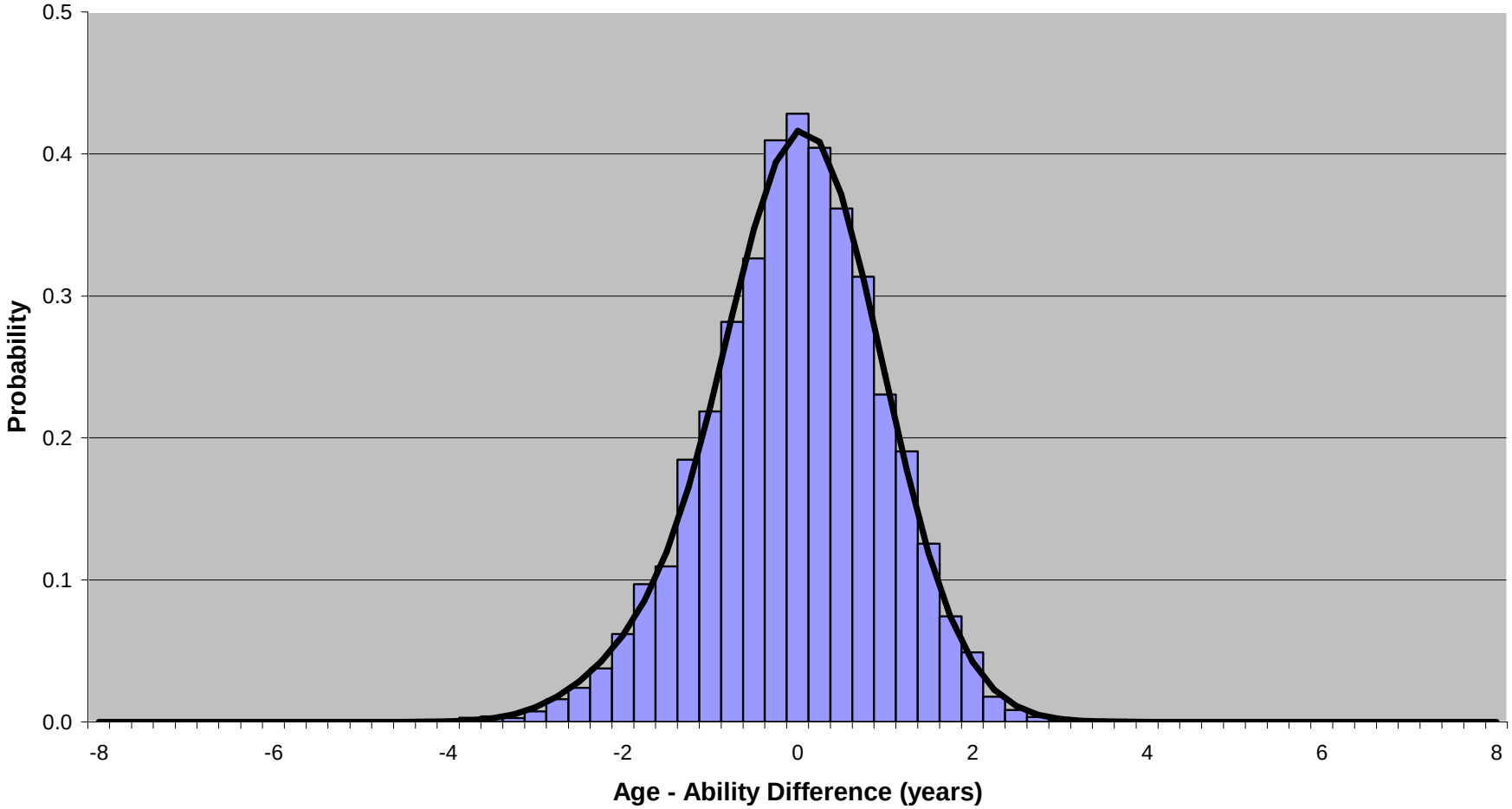


Figure 3.3.2.1: Binormal Model Plot for Boys' Mathematics Results in P5

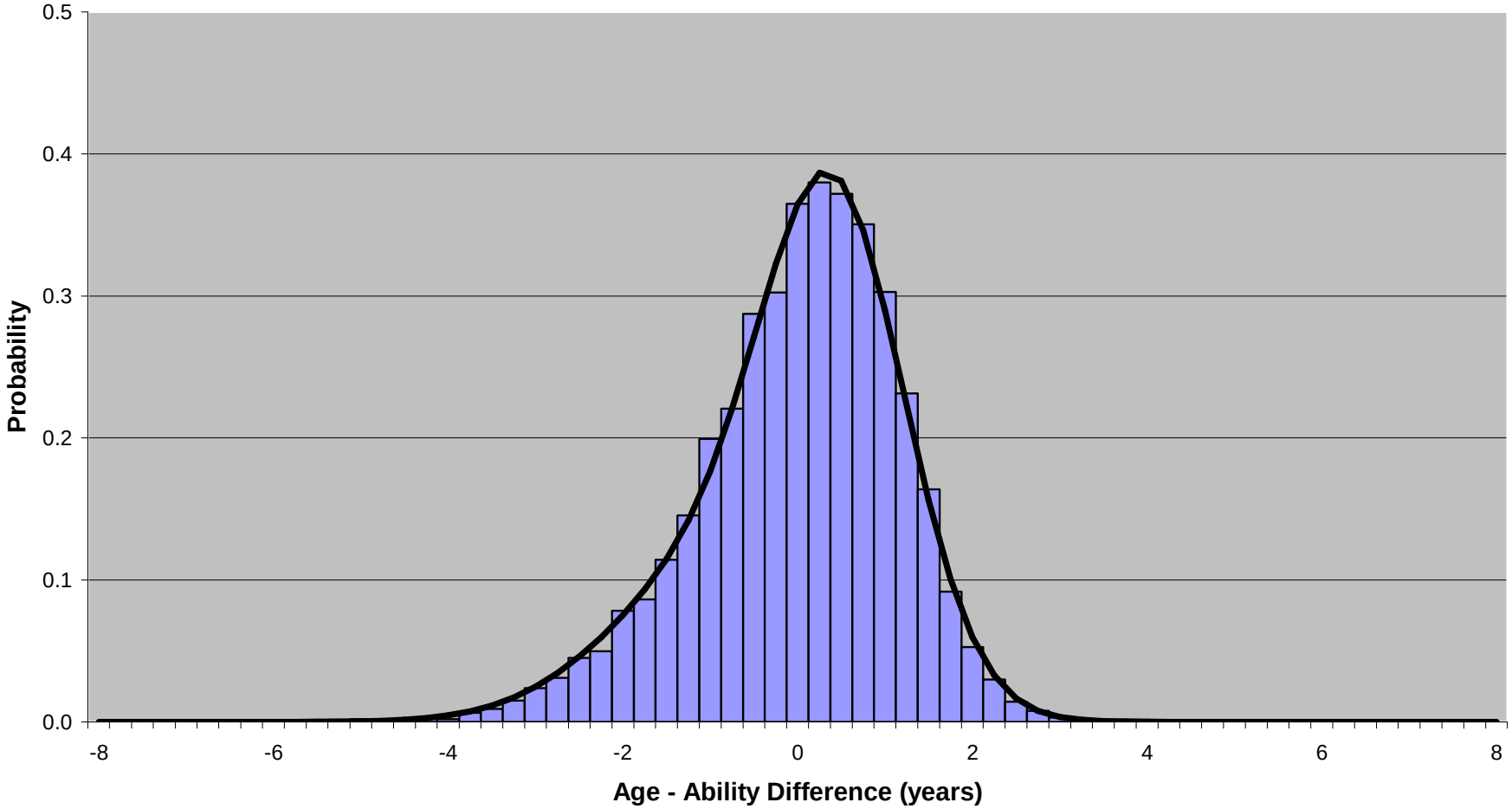


Figure 3.3.2.2: Binormal Model Plot for Girls' Mathematics Results in P5

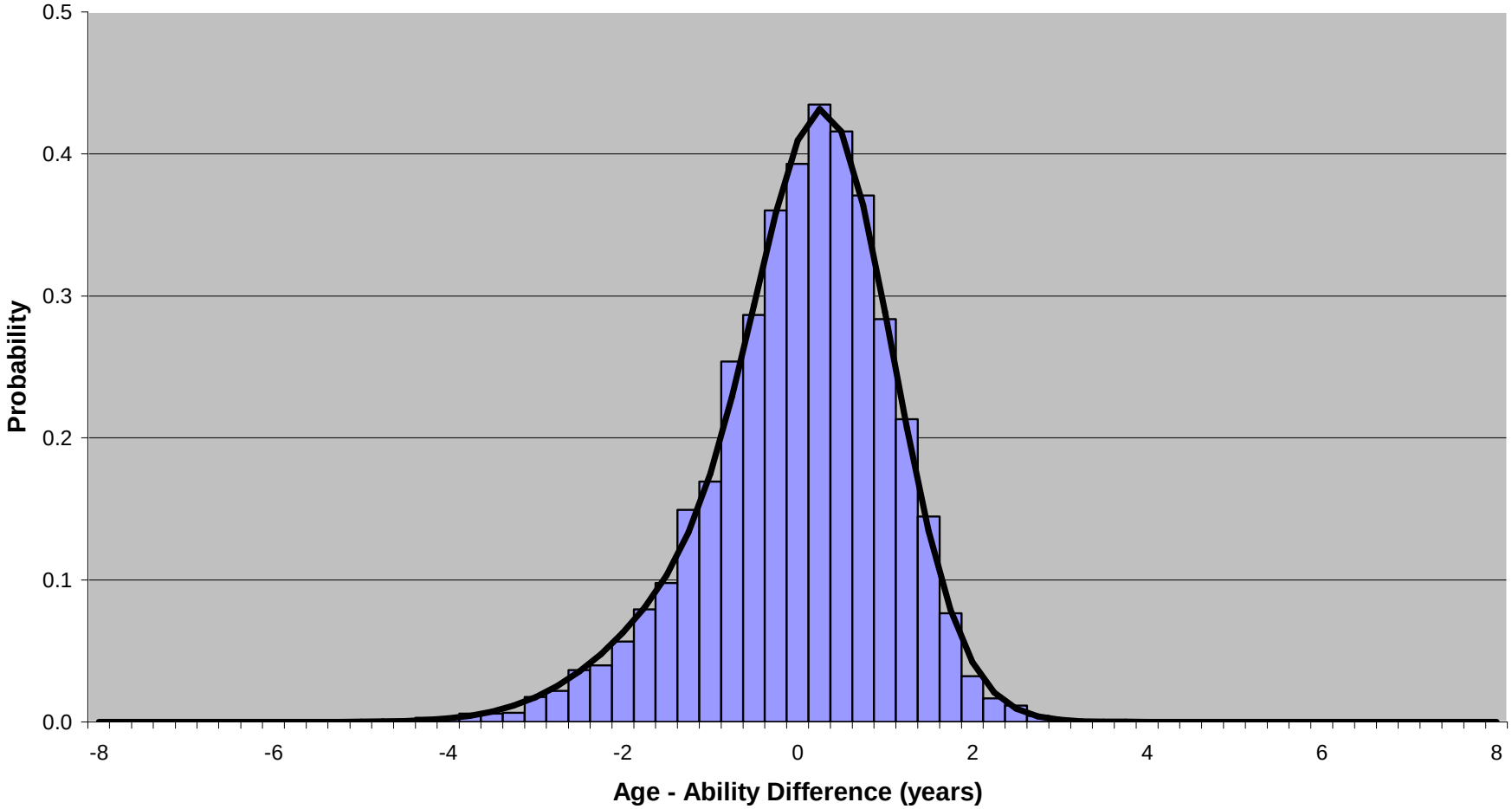


Figure 3.3.3.1: Binormal Model Plot for Boys' Mathematics Results in P6

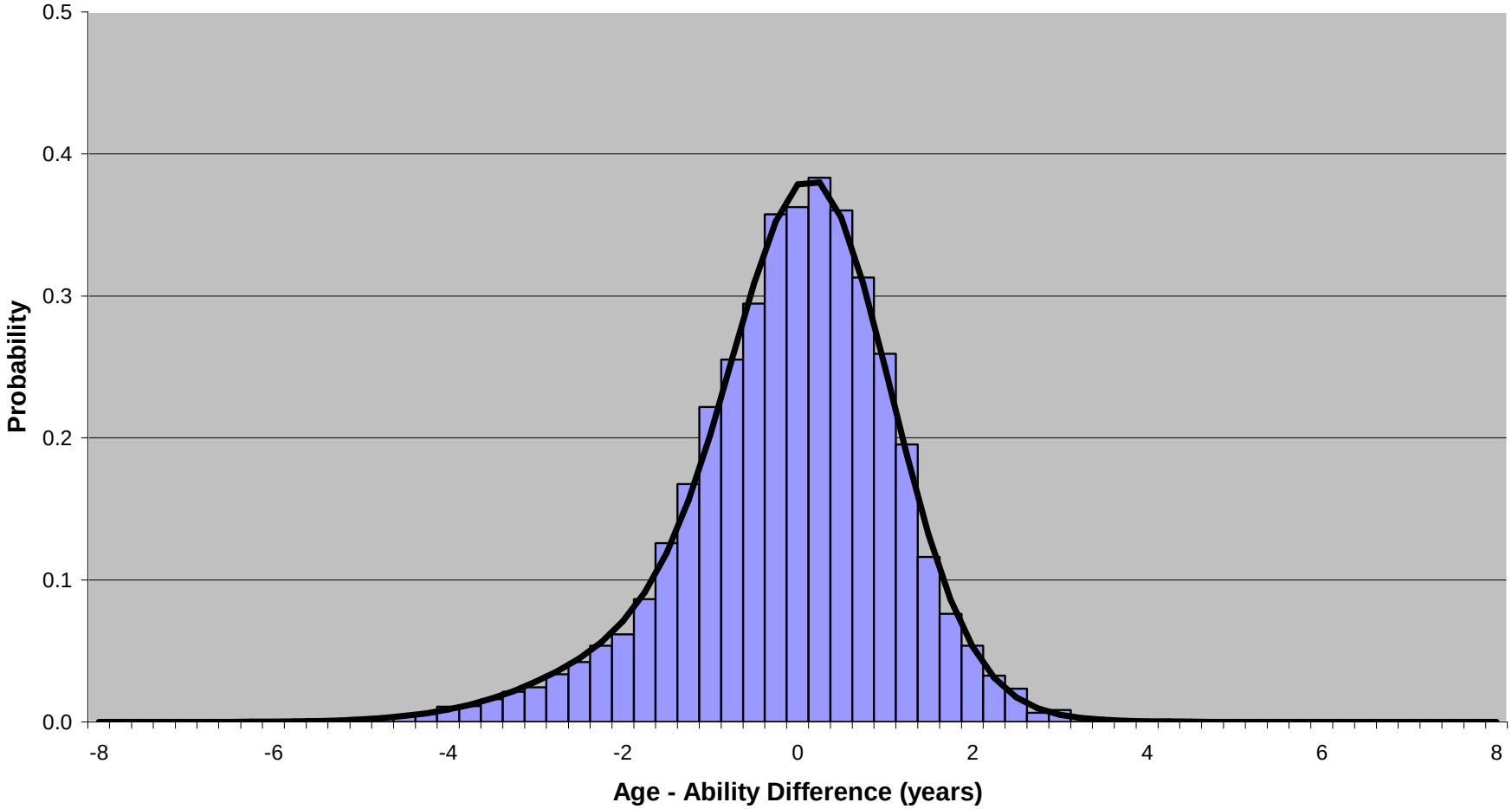


Figure 3.3.3.2: Binormal Model Plot for Girls' Mathematics Results in P6

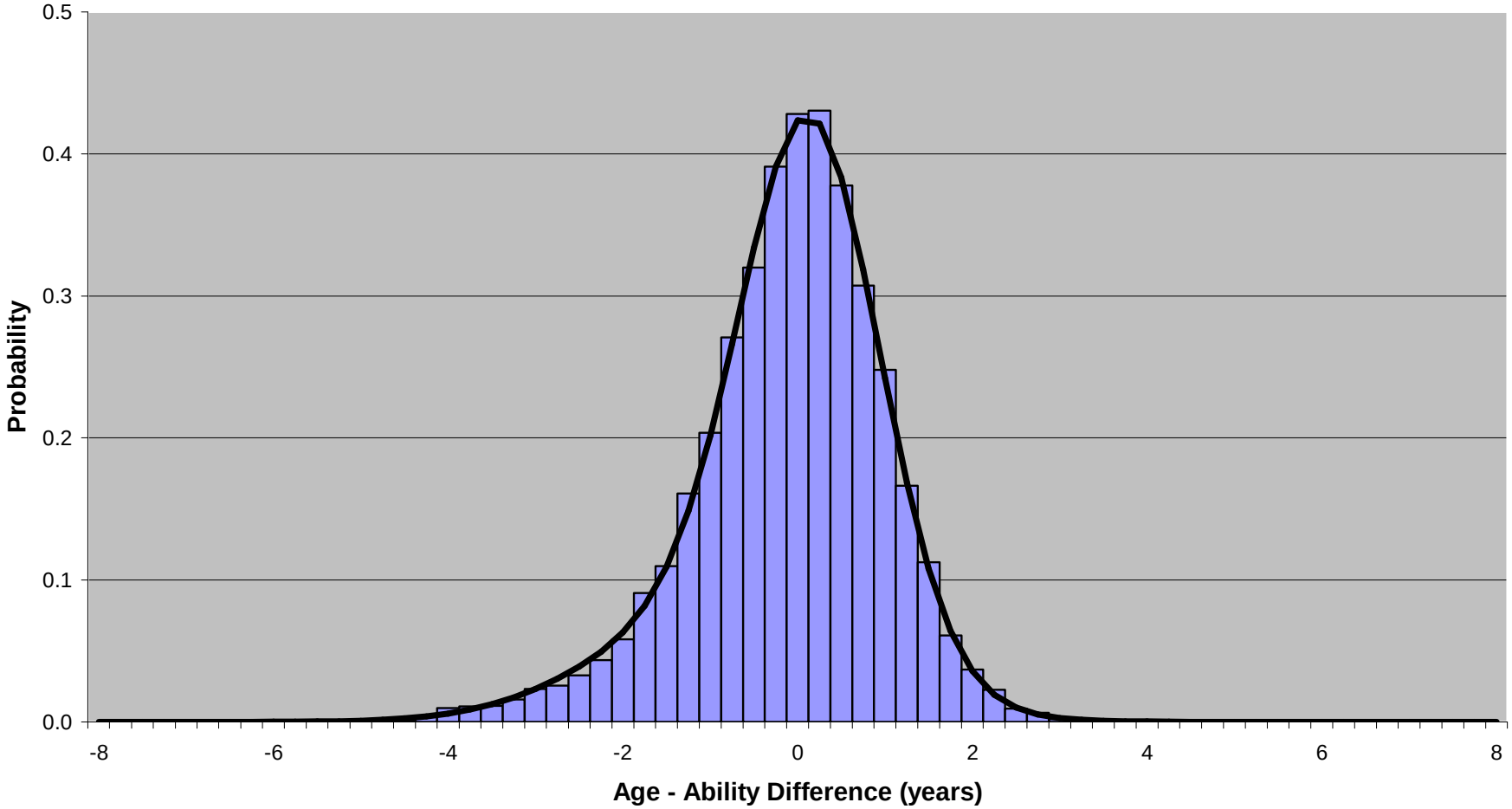


Figure 3.3.4.1: Binormal Model Plot for Boys' Mathematics Results in P7

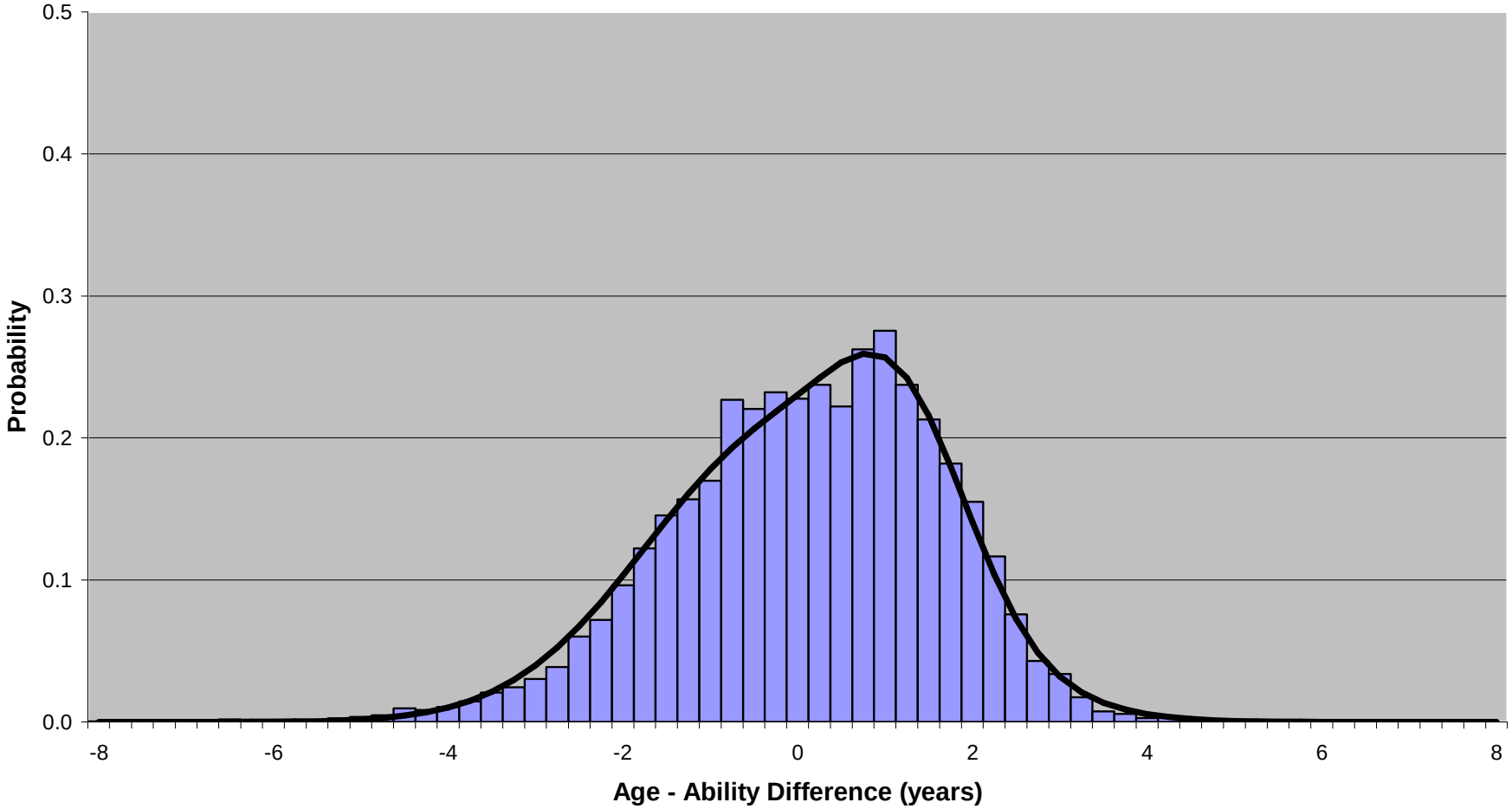


Figure 3.3.4.2: Binormal Model Plot for Girls' Mathematics Results in P7

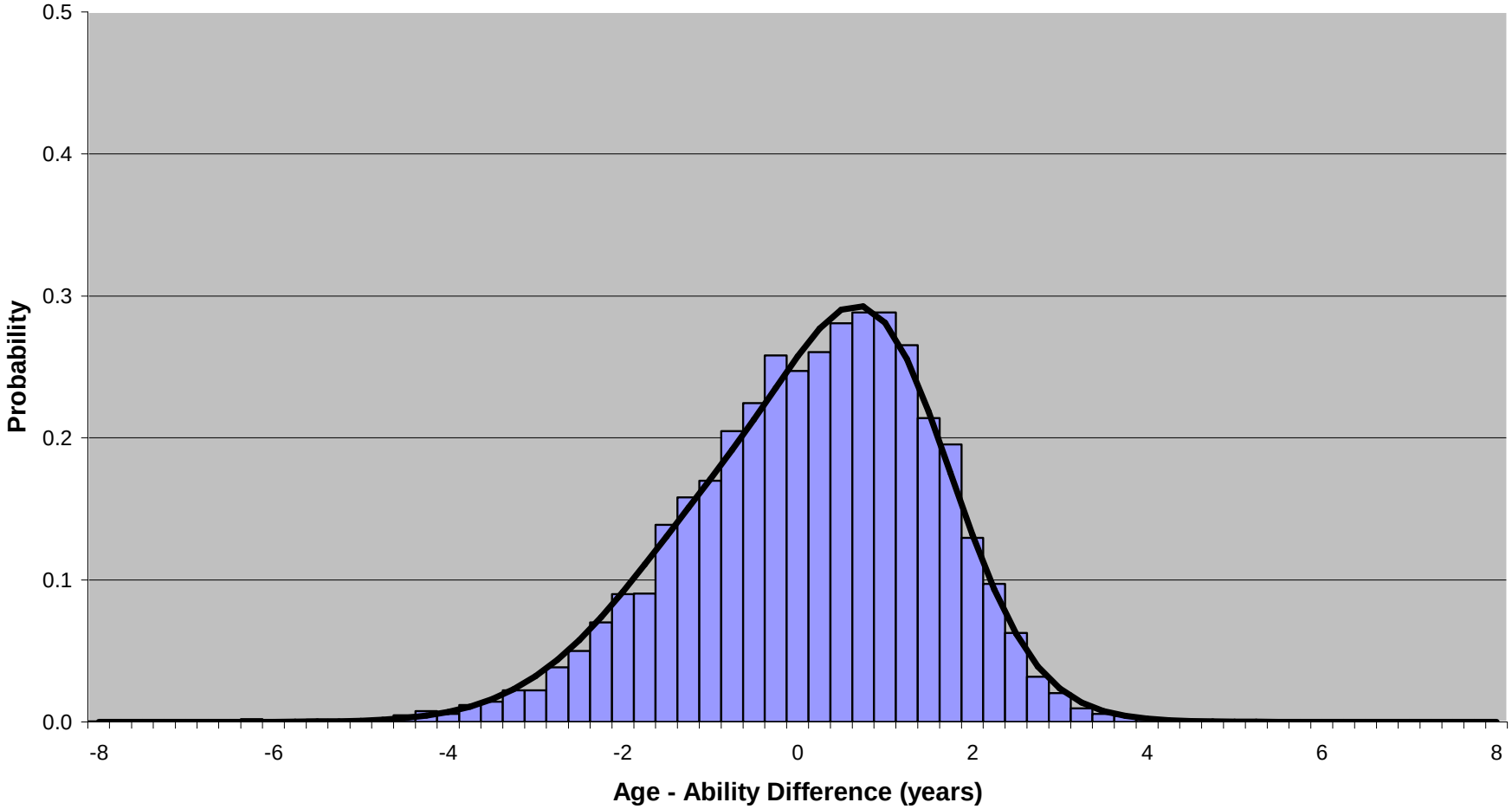


Figure 3.4.1.1: Binormal Model Plot for Boys' Arithmetic Results in P4

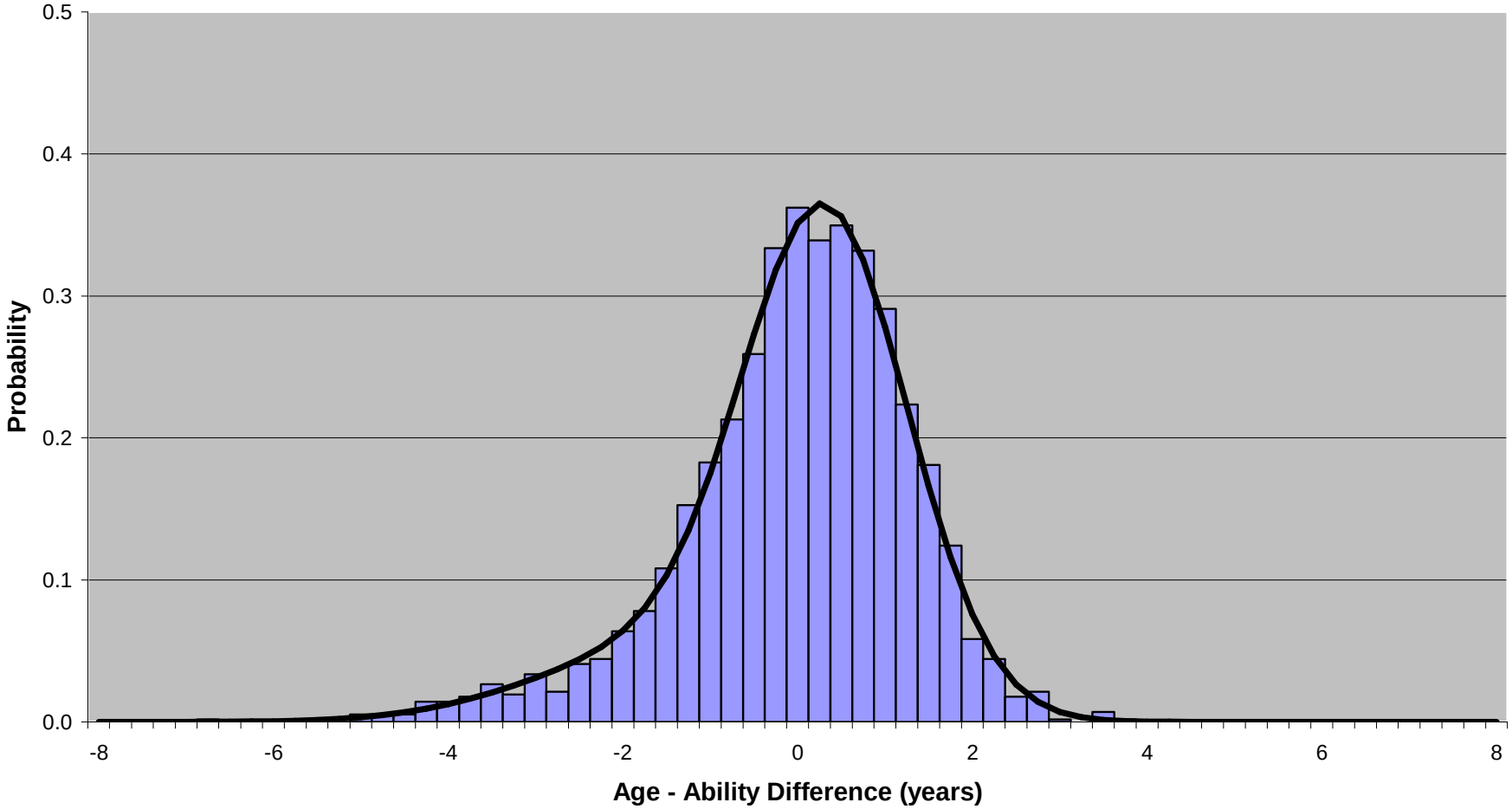


Figure 3.4.1.2: Binormal Model Plot for Girls' Arithmetic Results in P4

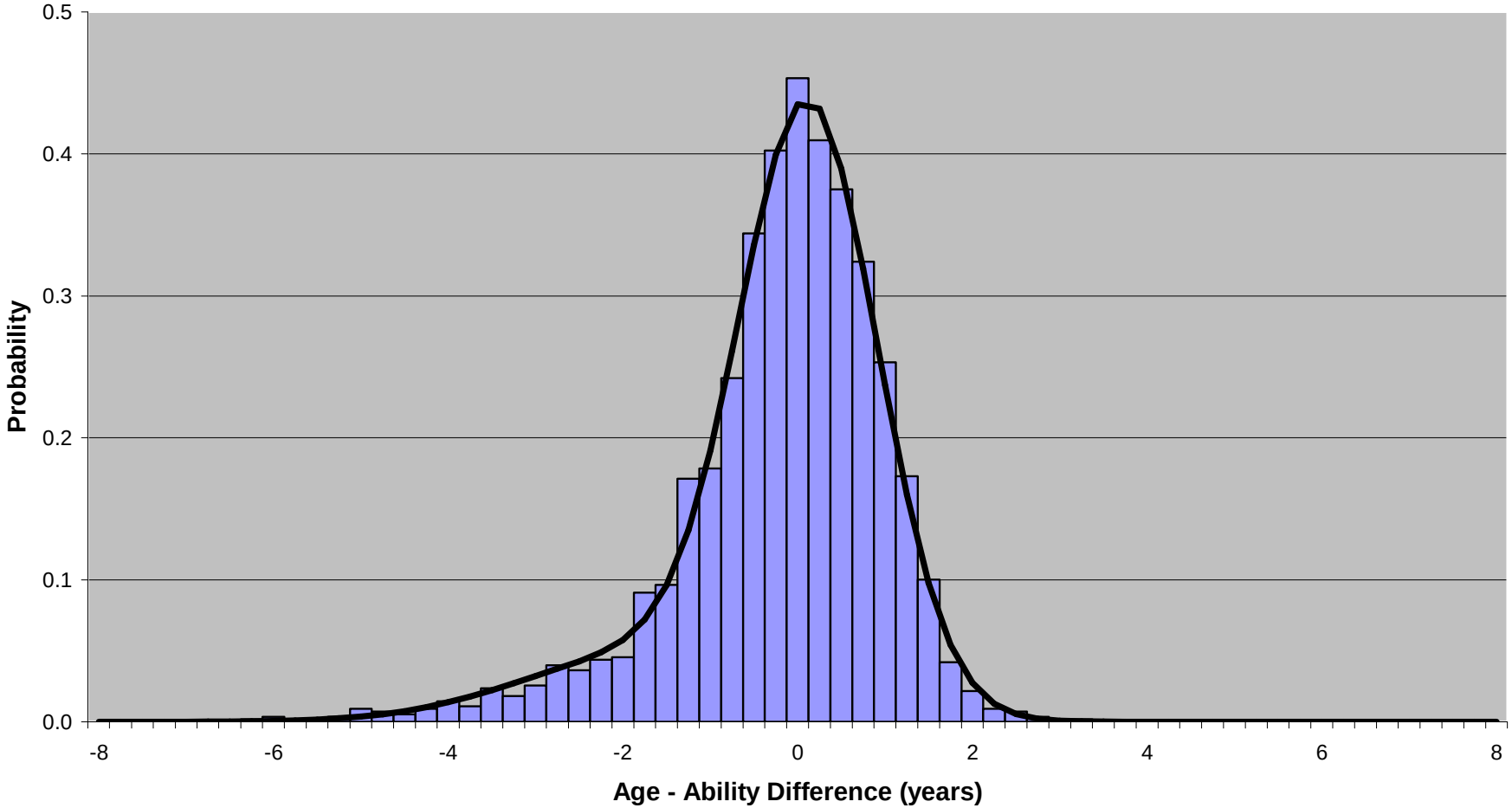


Figure 3.4.2.1: Binormal Model Plot for Boys' Arithmetic Results in P5

Figure 3.4.2.2: Binormal Model Plot for Girls' Arithmetic Results in P5

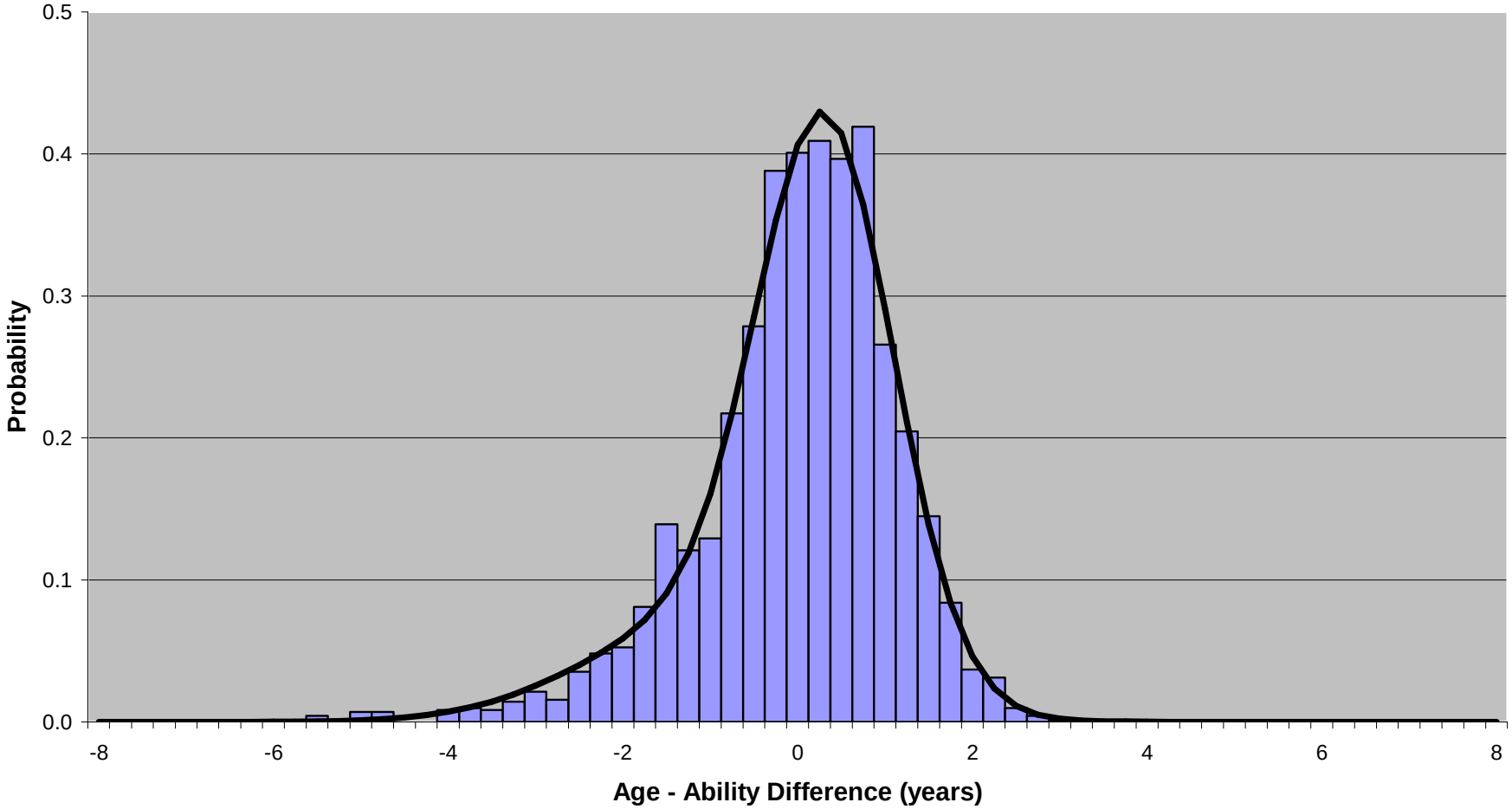


Figure 3.4.3.1: Binormal Model Plot for Boys' Arithmetic Results in P6

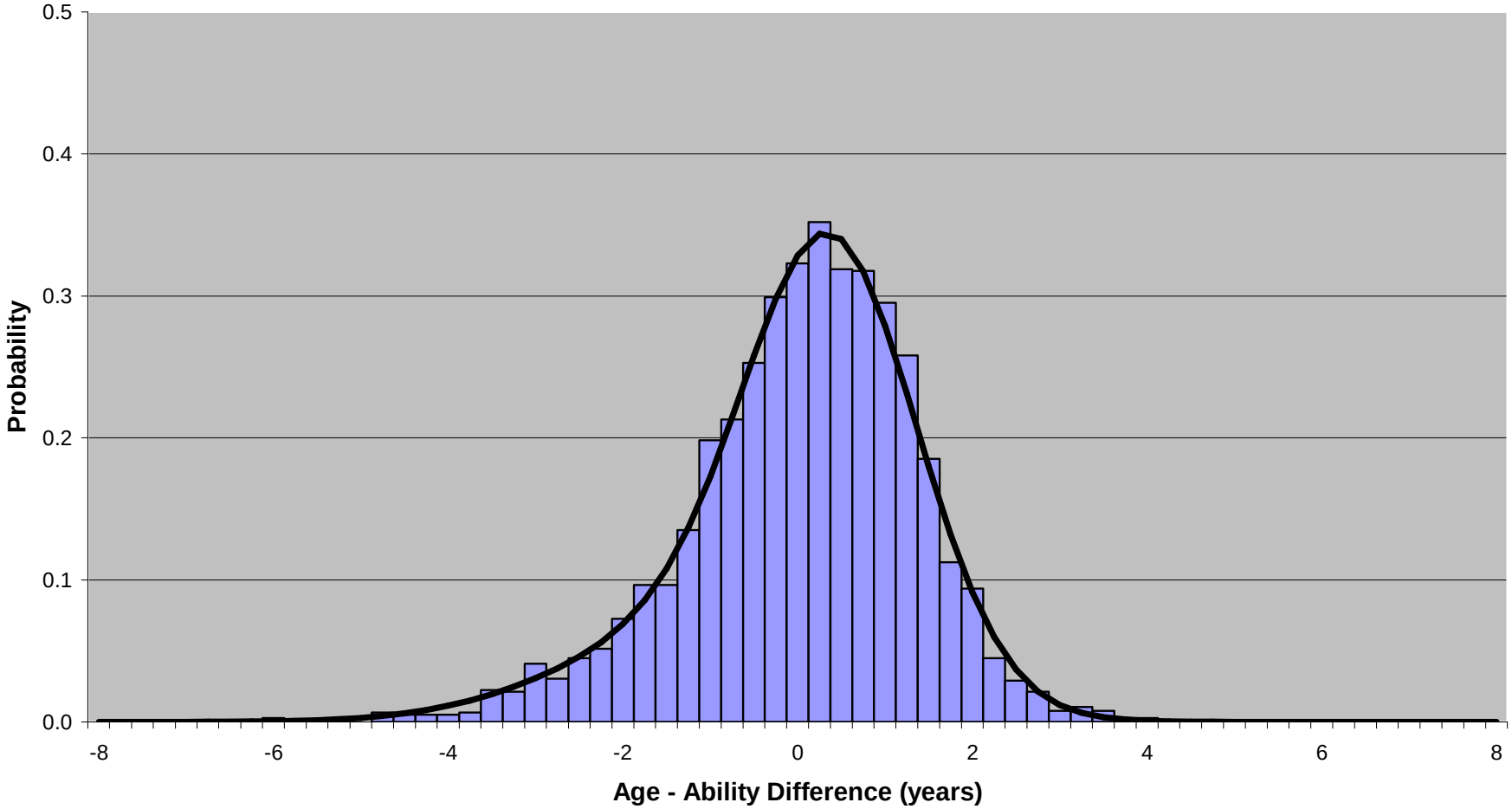


Figure 3.4.3.2: Binormal Model Plot for Girls' Arithmetic Results in P6

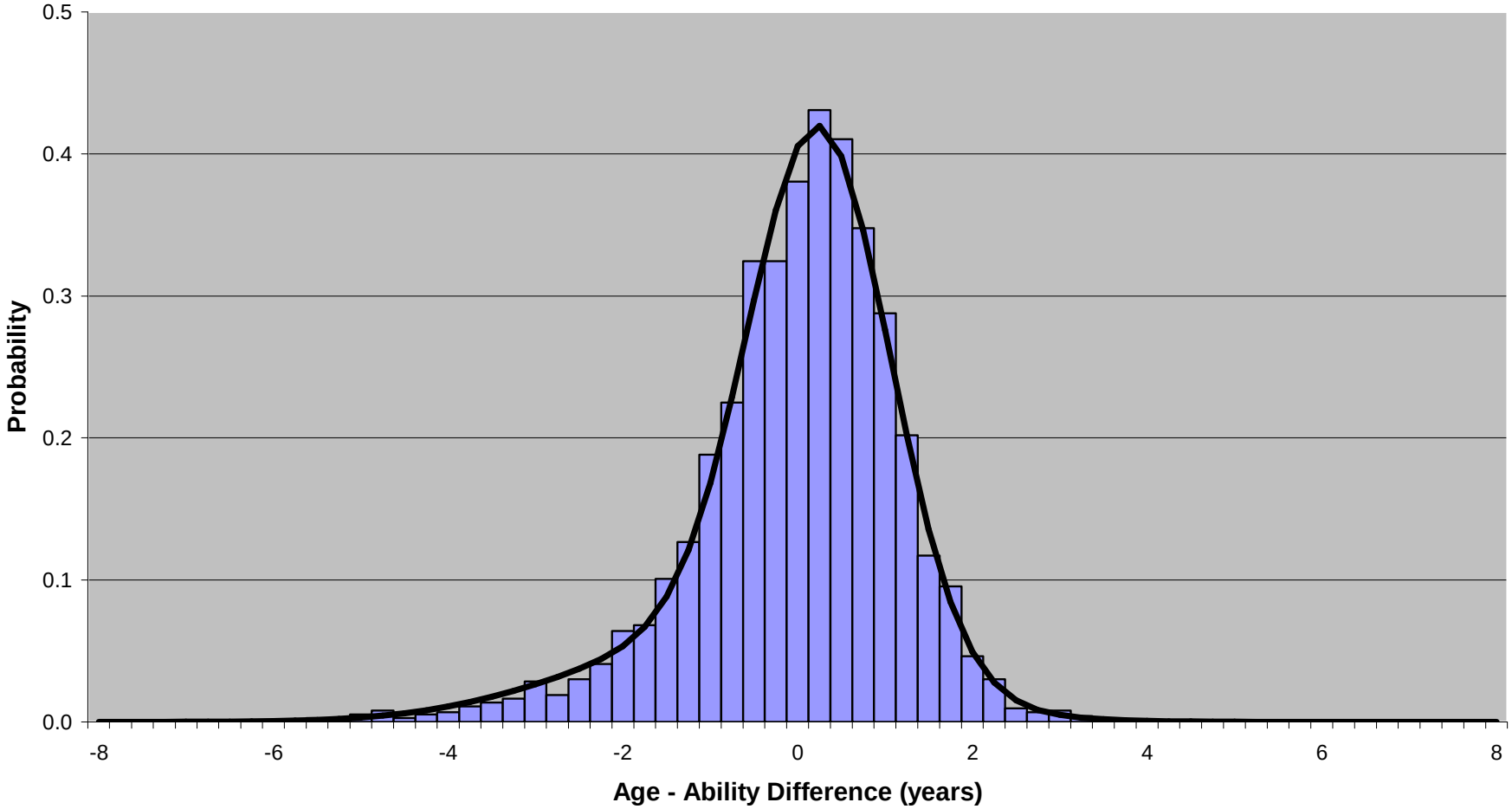


Figure 3.4.4.1: Binormal Model Plot for Boys' Arithmetic Results in P7

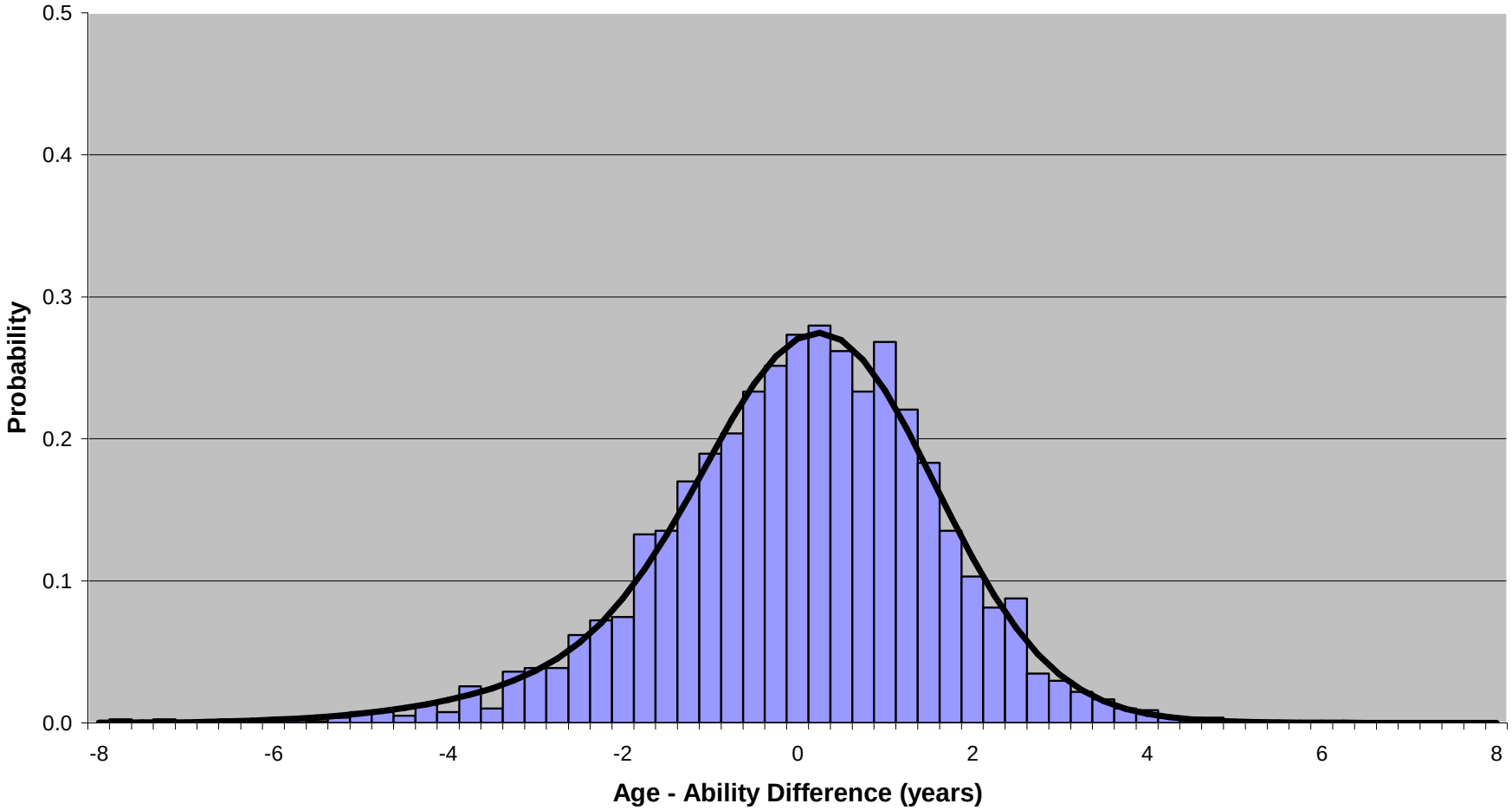


Figure 3.4.4.2: Binormal Model Plot for Girls' Arithmetic Results in P7

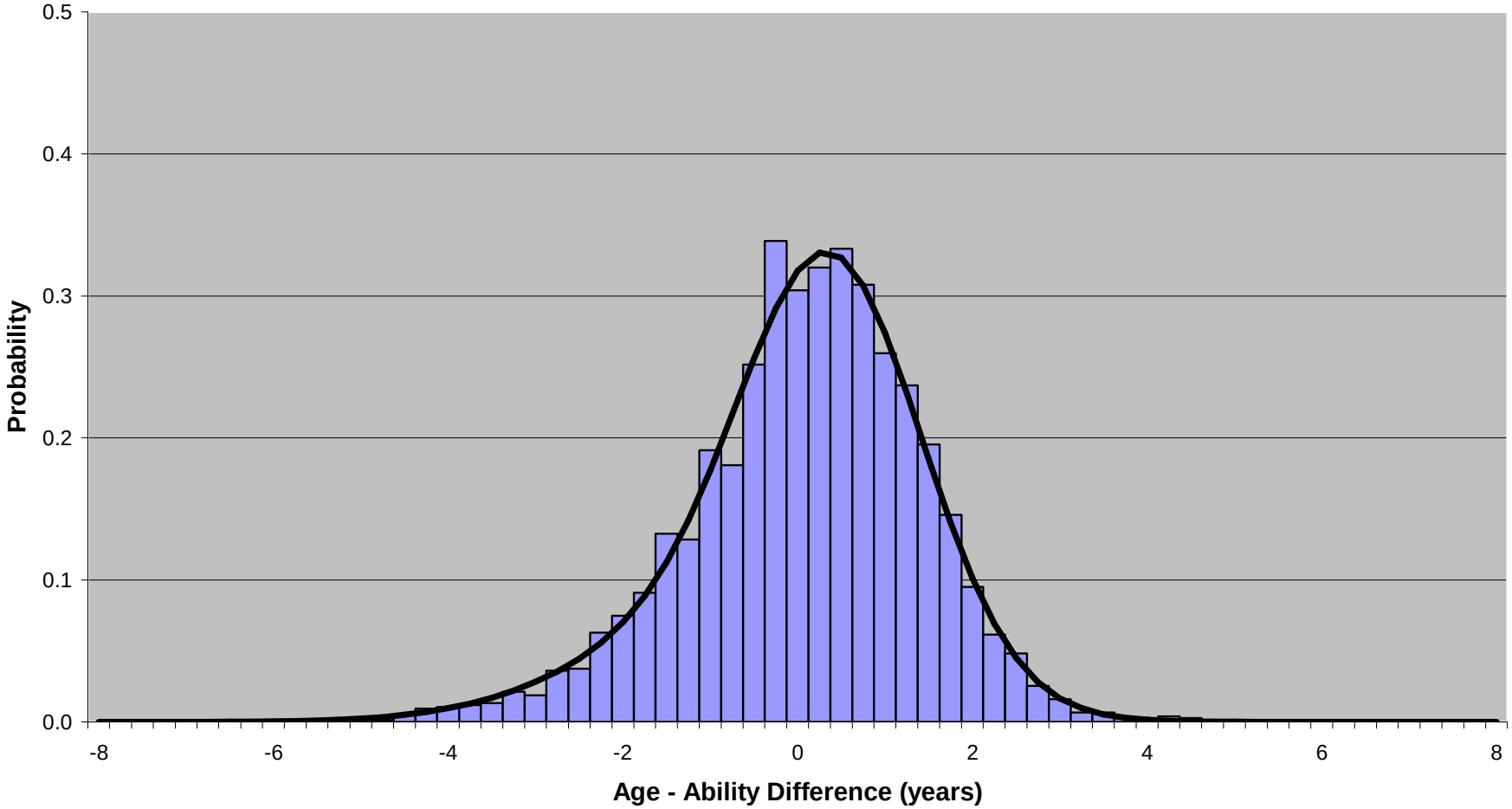


Figure 4.1.1.1: Binormal Subpopulation Plot for Boys' Picture Vocabulary Results in P4

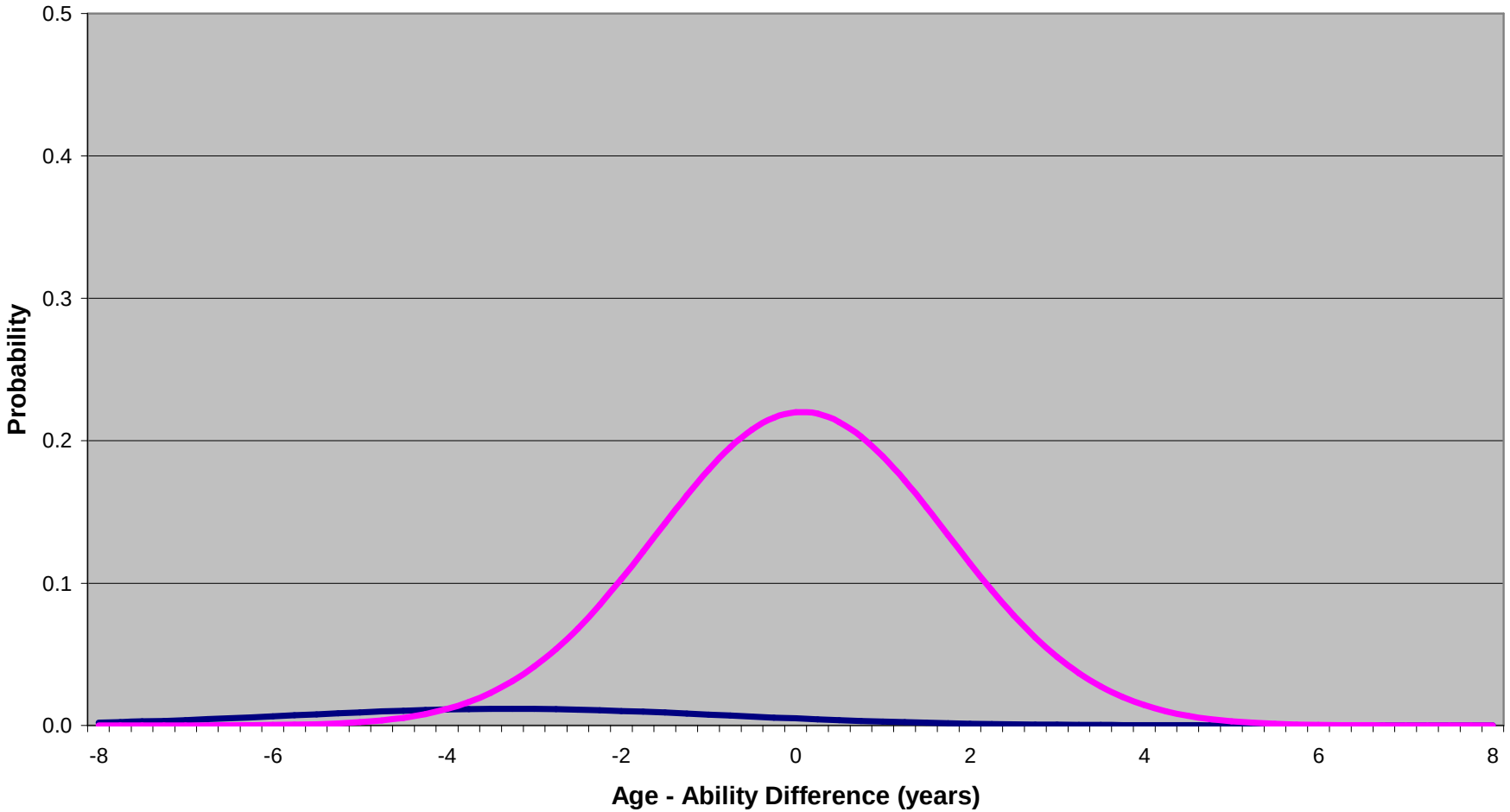


Figure 4.1.1.2: Binormal Subpopulation Plot for Girls' Picture Vocabulary Results in P4

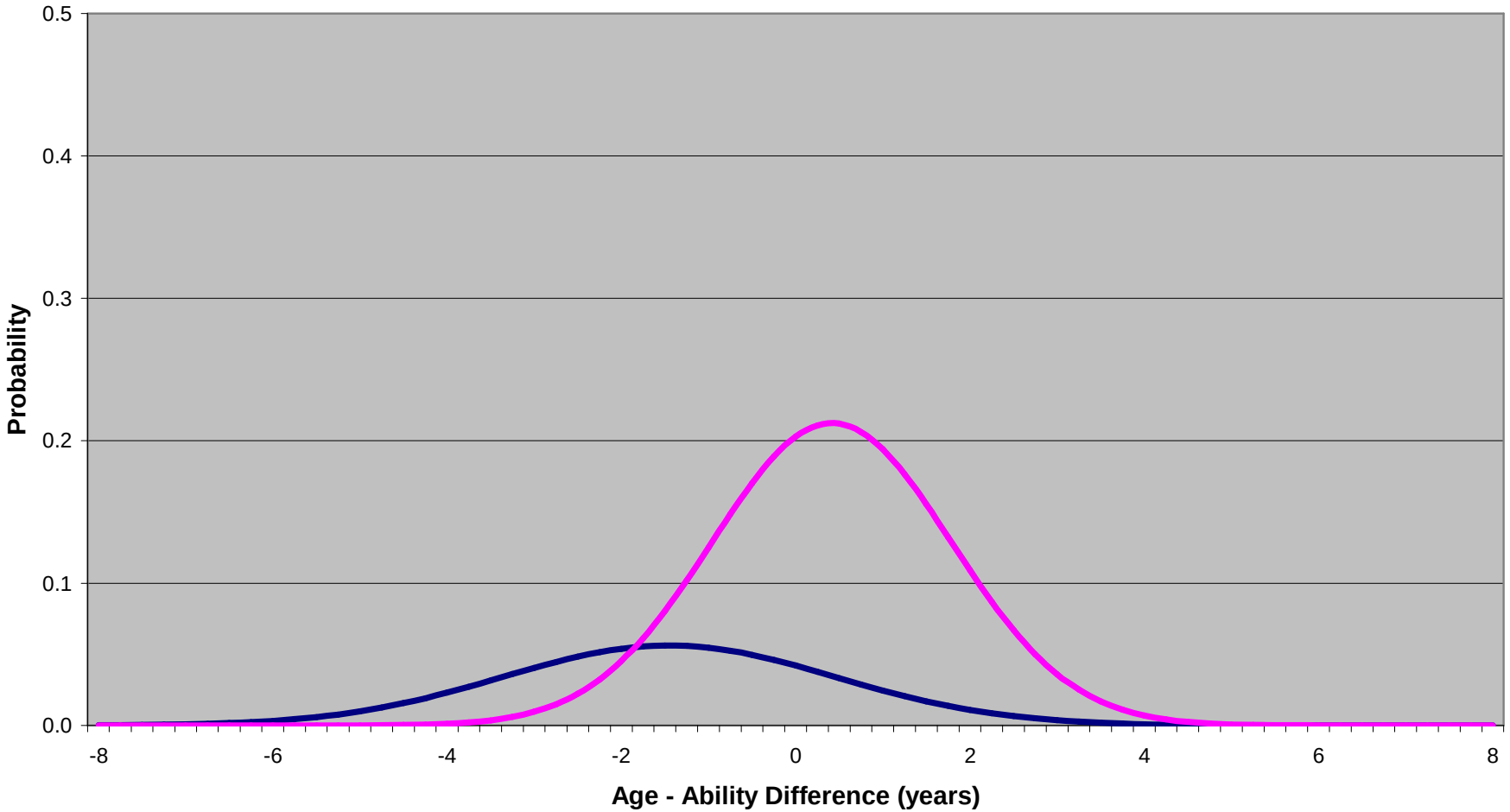


Figure 4.1.2.1: Binormal Subpopulation Plot for Boys' Picture Vocabulary Results in P5

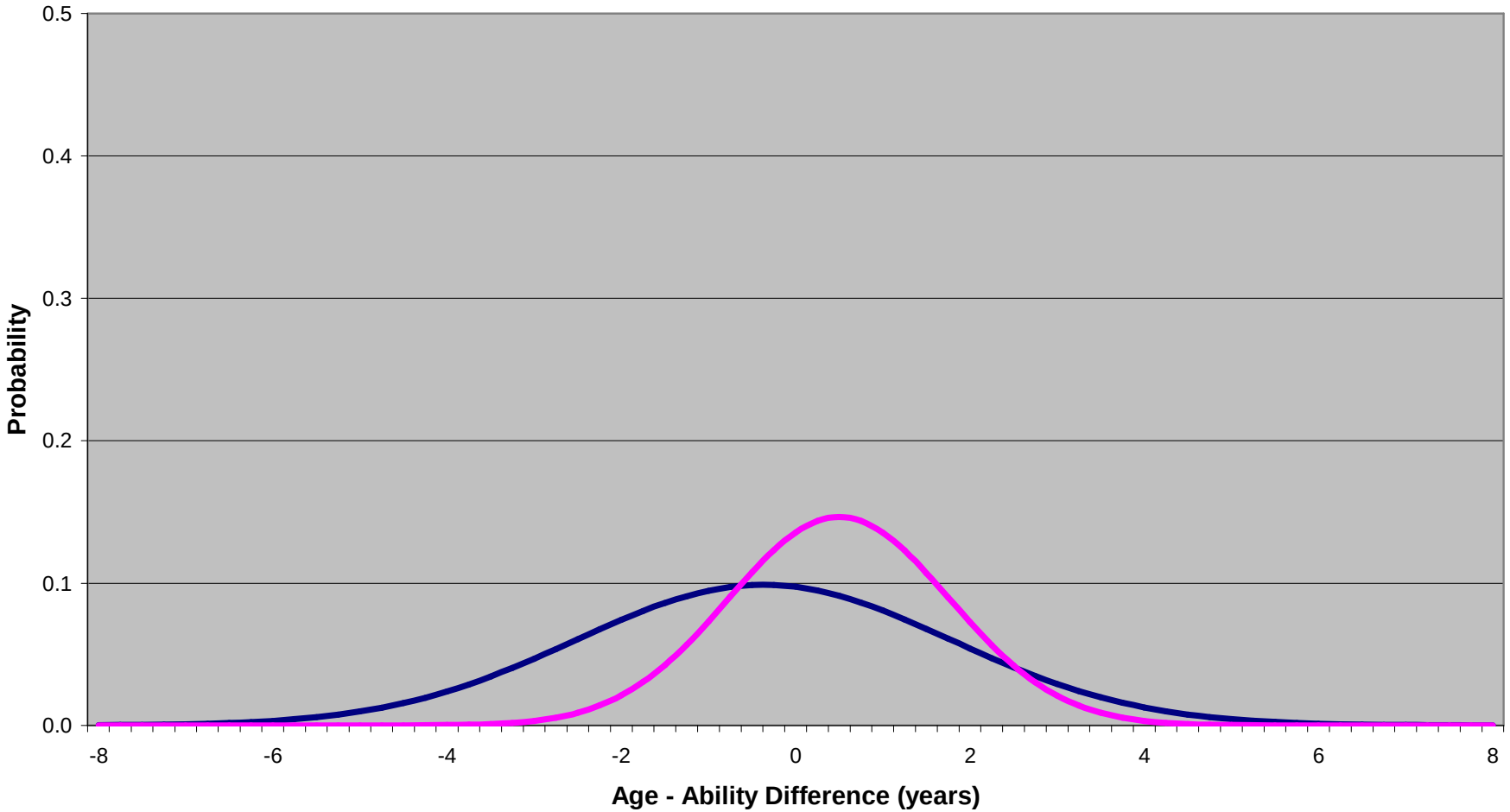


Figure 4.1.2.2: Binormal Subpopulation Plot for Girls' Picture Vocabulary Results in P5

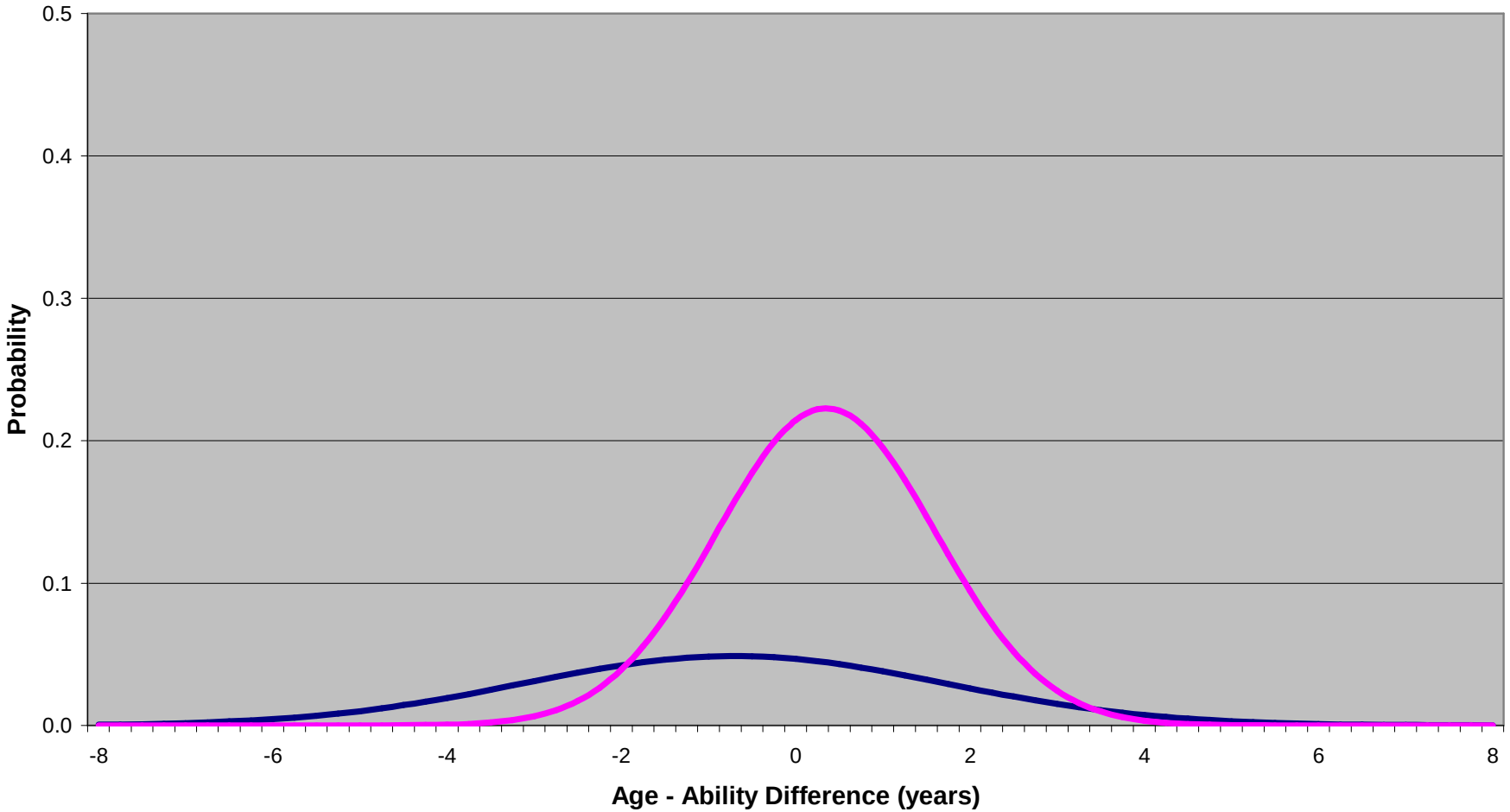


Figure 4.1.3.1: Binormal Subpopulation Plot for Boys' Picture Vocabulary Results in P6

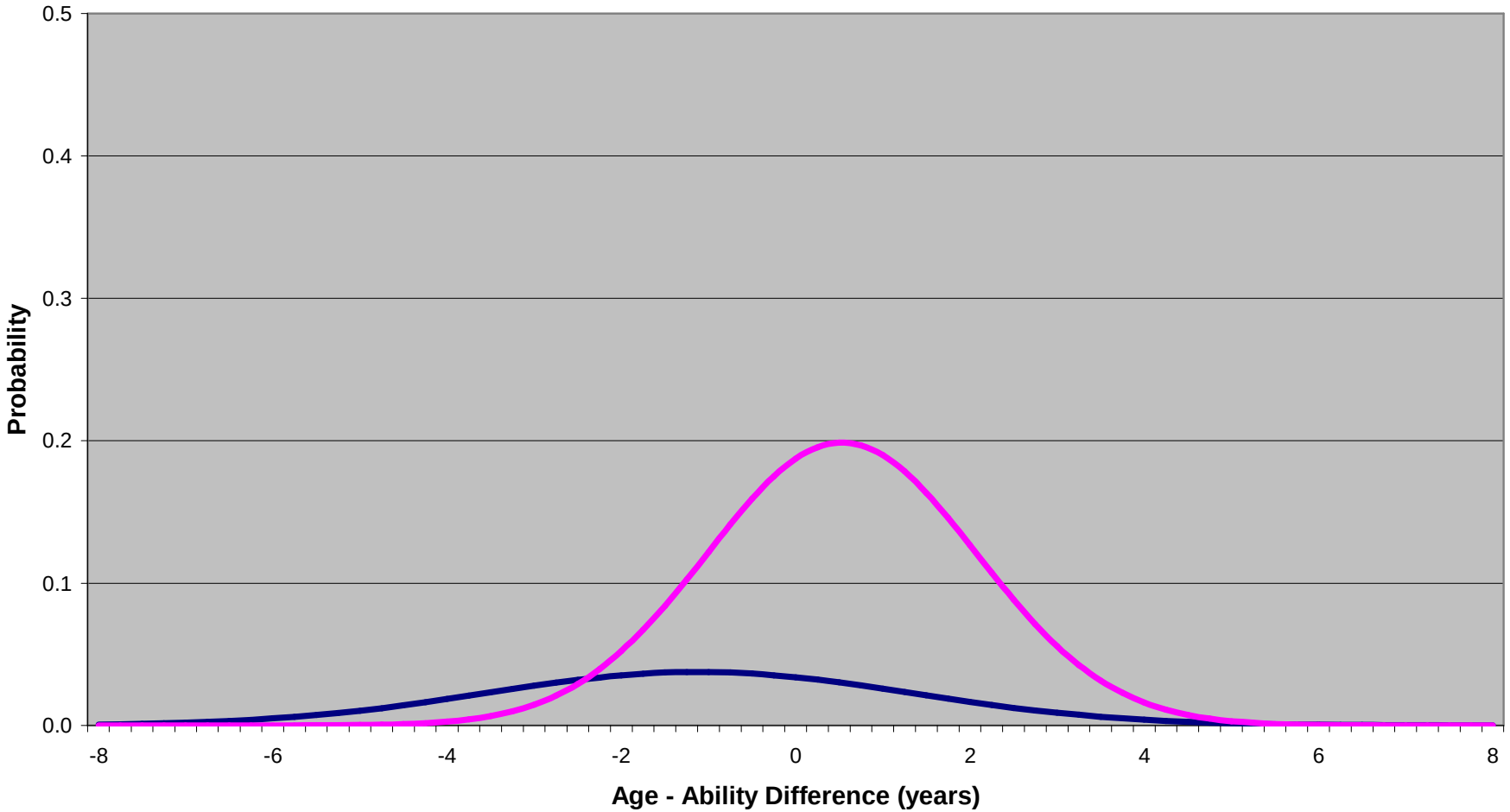


Figure 4.1.3.2: Binormal Subpopulation Plot for Girls' Picture Vocabulary Results in P6

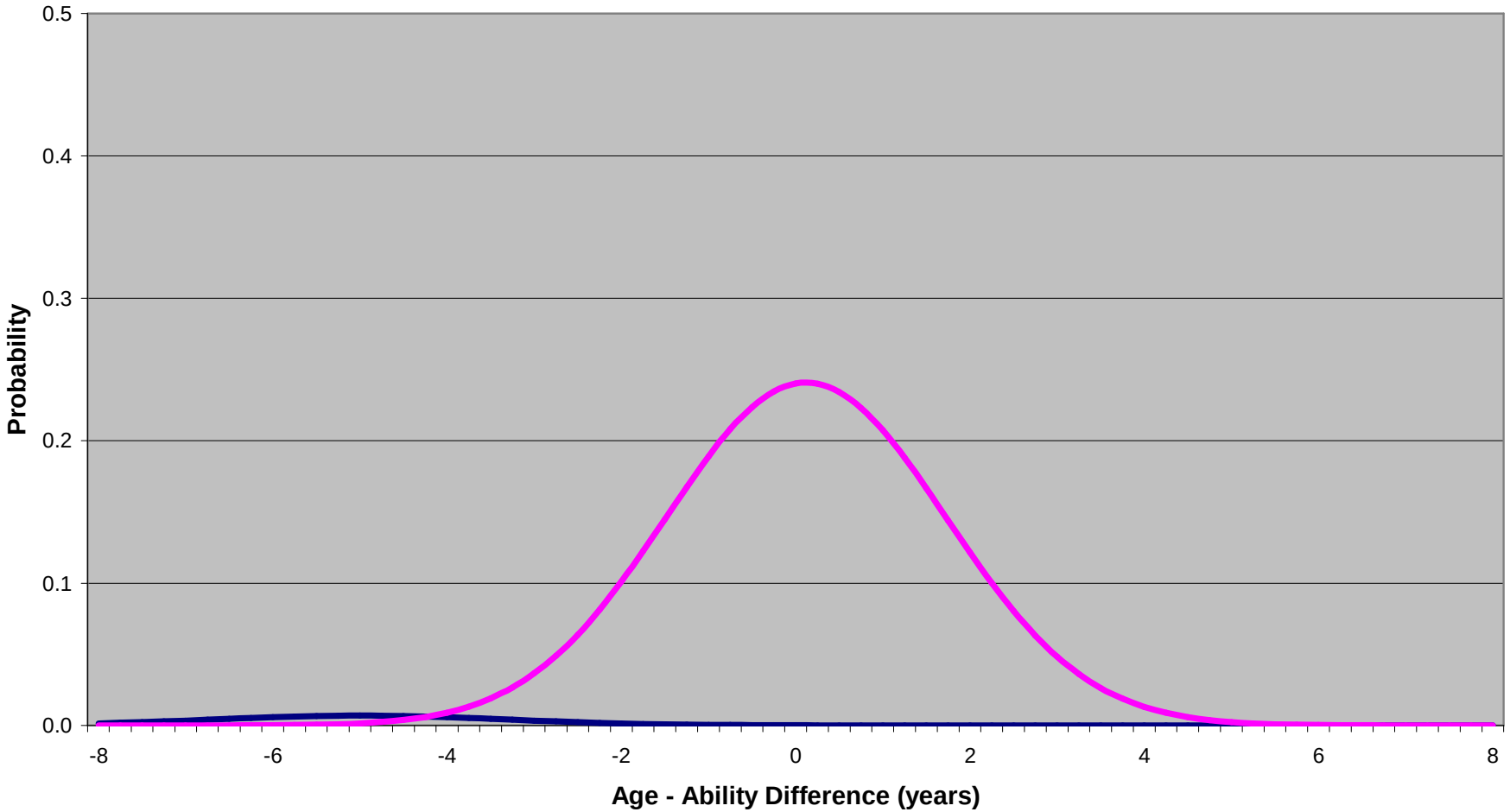


Figure 4.1.4.1: Binormal Subpopulation Plot for Boys' Picture Vocabulary Results in P7

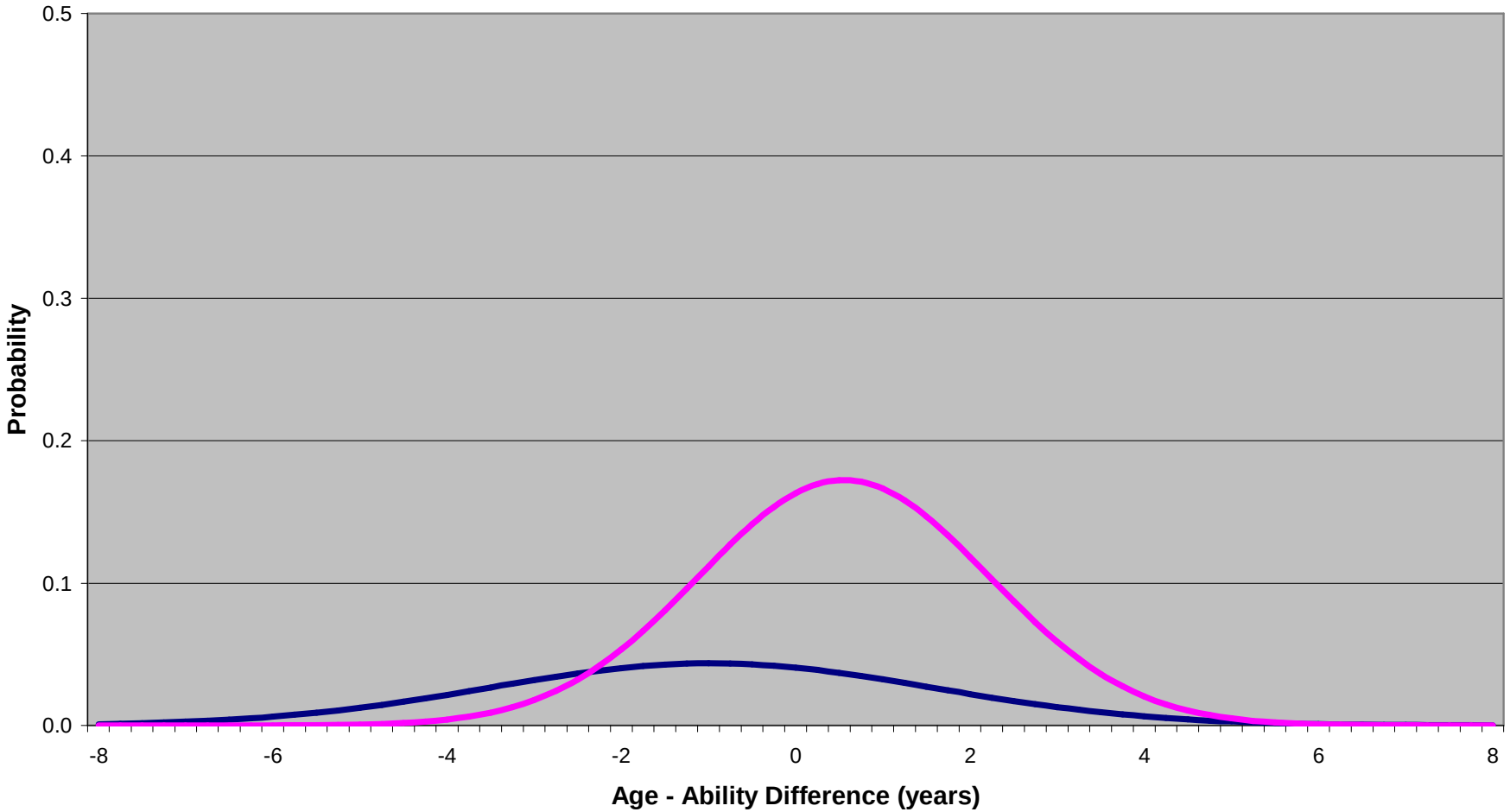


Figure 4.1.4.2: Binormal Subpopulation Plot for Girls' Picture Vocabulary Results in P7

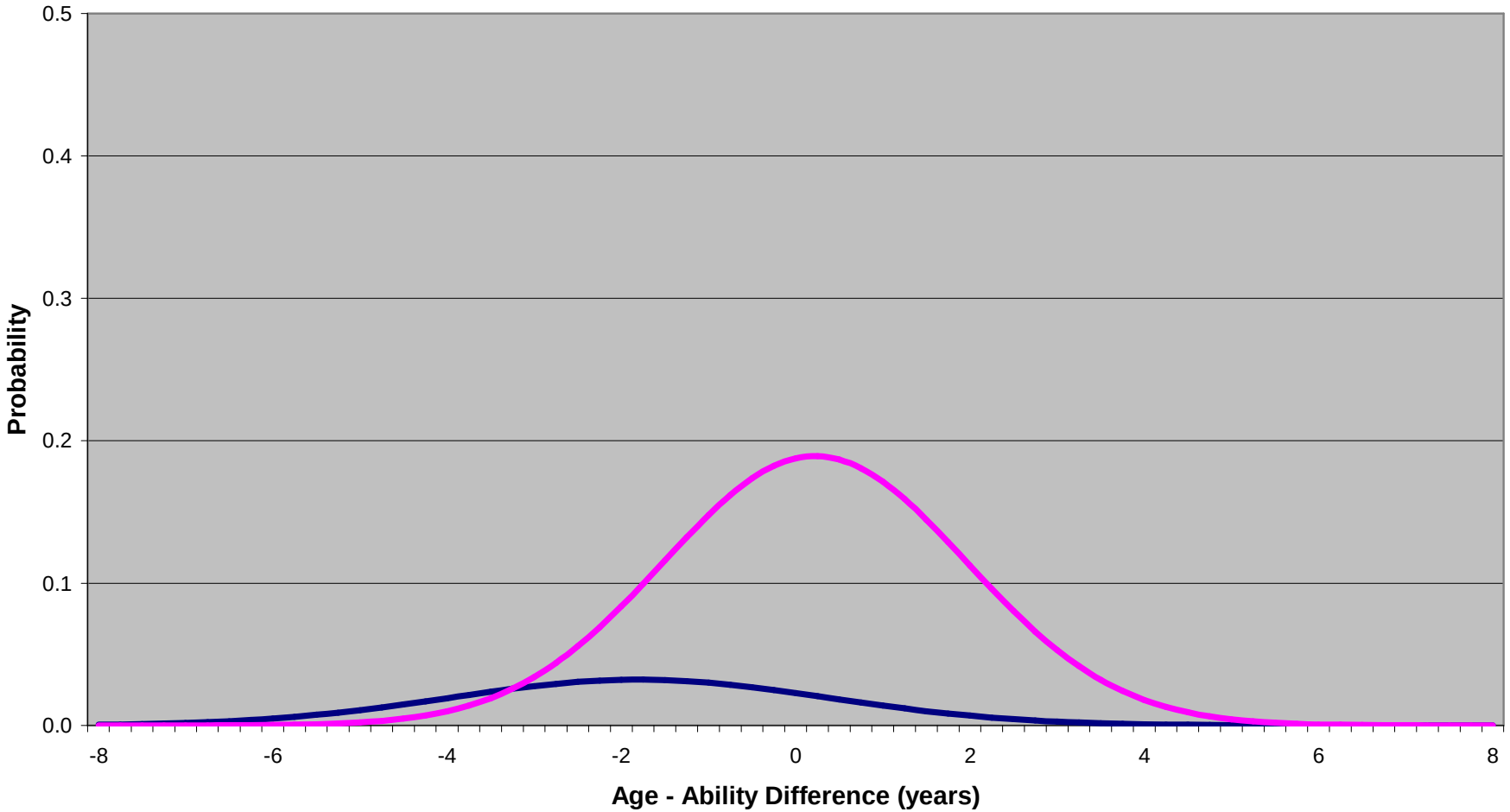


Figure 4.2.1.1: Binormal Subpopulation Plot for Boys' Reading Results in P4

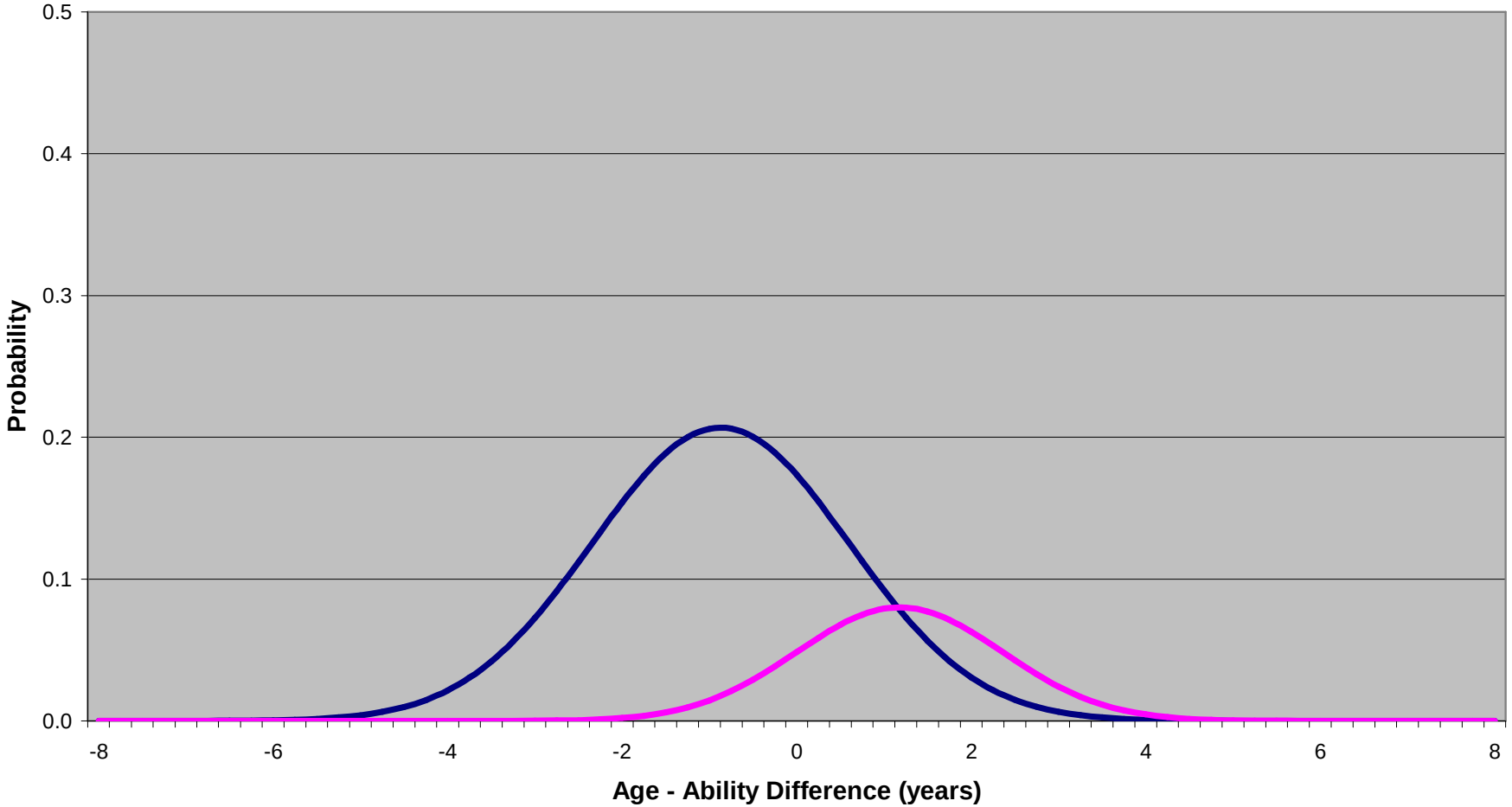


Figure 4.2.1.2: Binormal Subpopulation Plot for Girls' Reading Results in P4

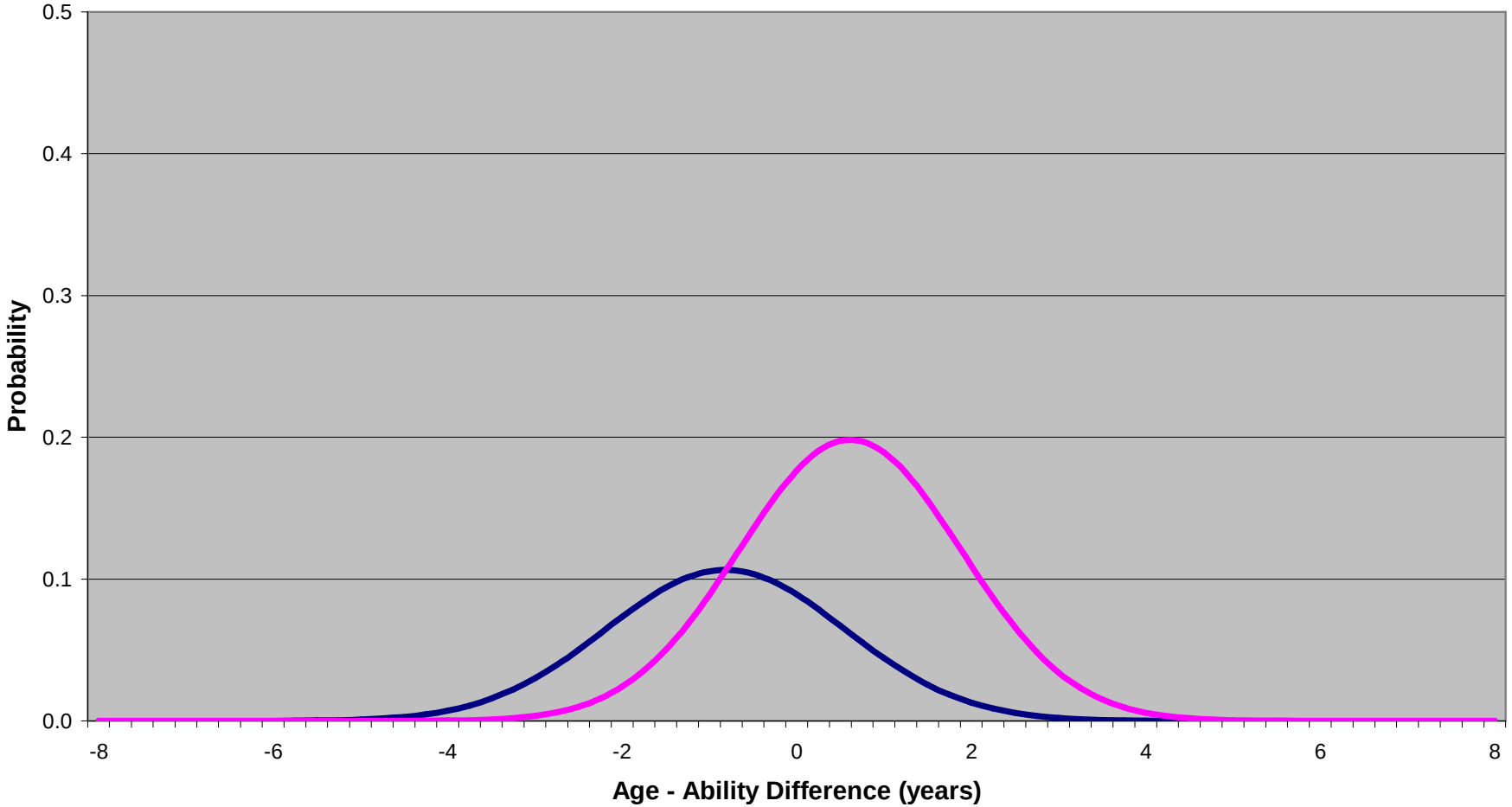


Figure 4.2.2.1: Binormal Subpopulation Plot for Boys' Reading Results in P5

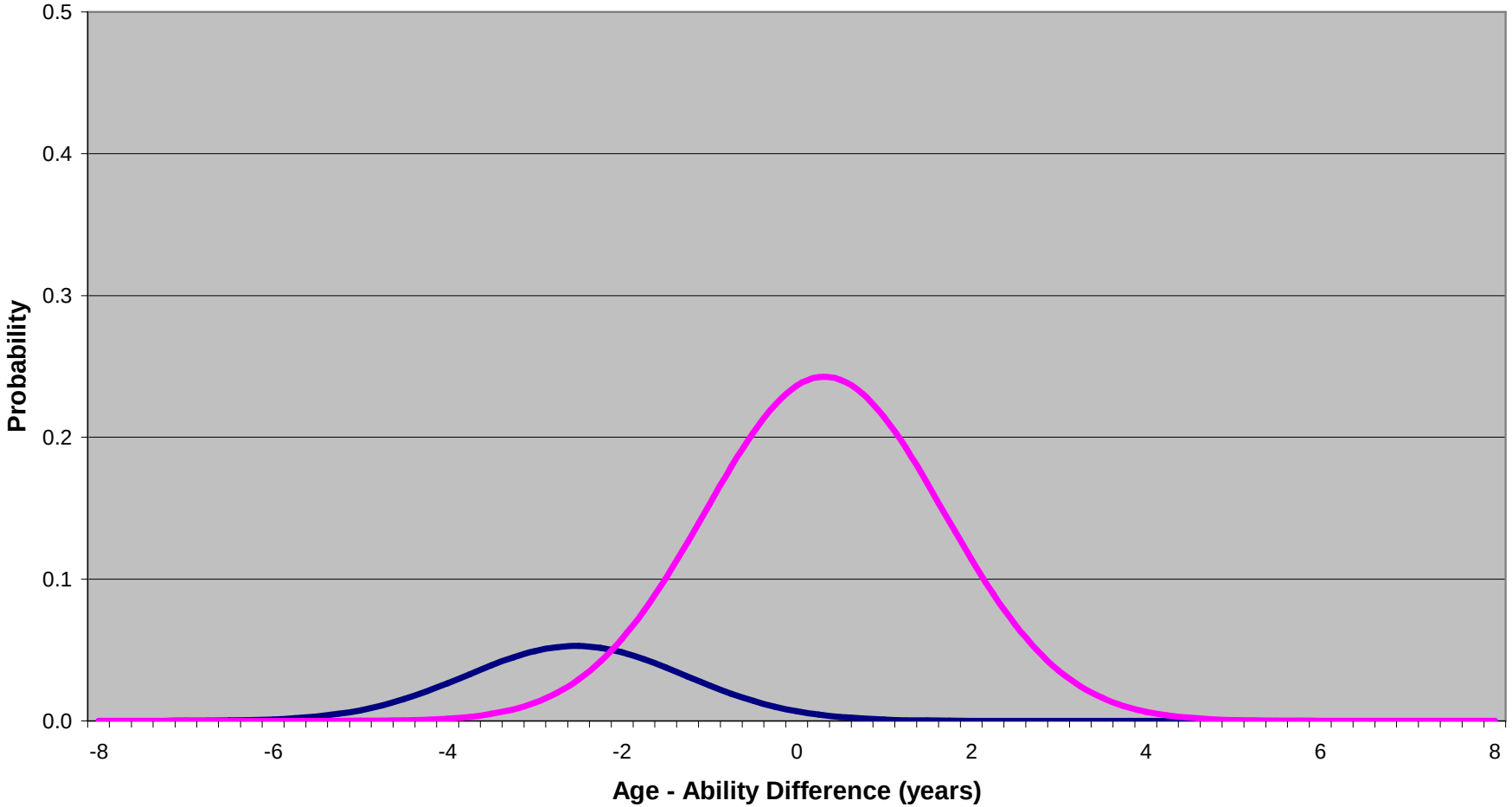


Figure 4.2.2.2: Binormal Subpopulation Plot for Girls' Reading Results in P5

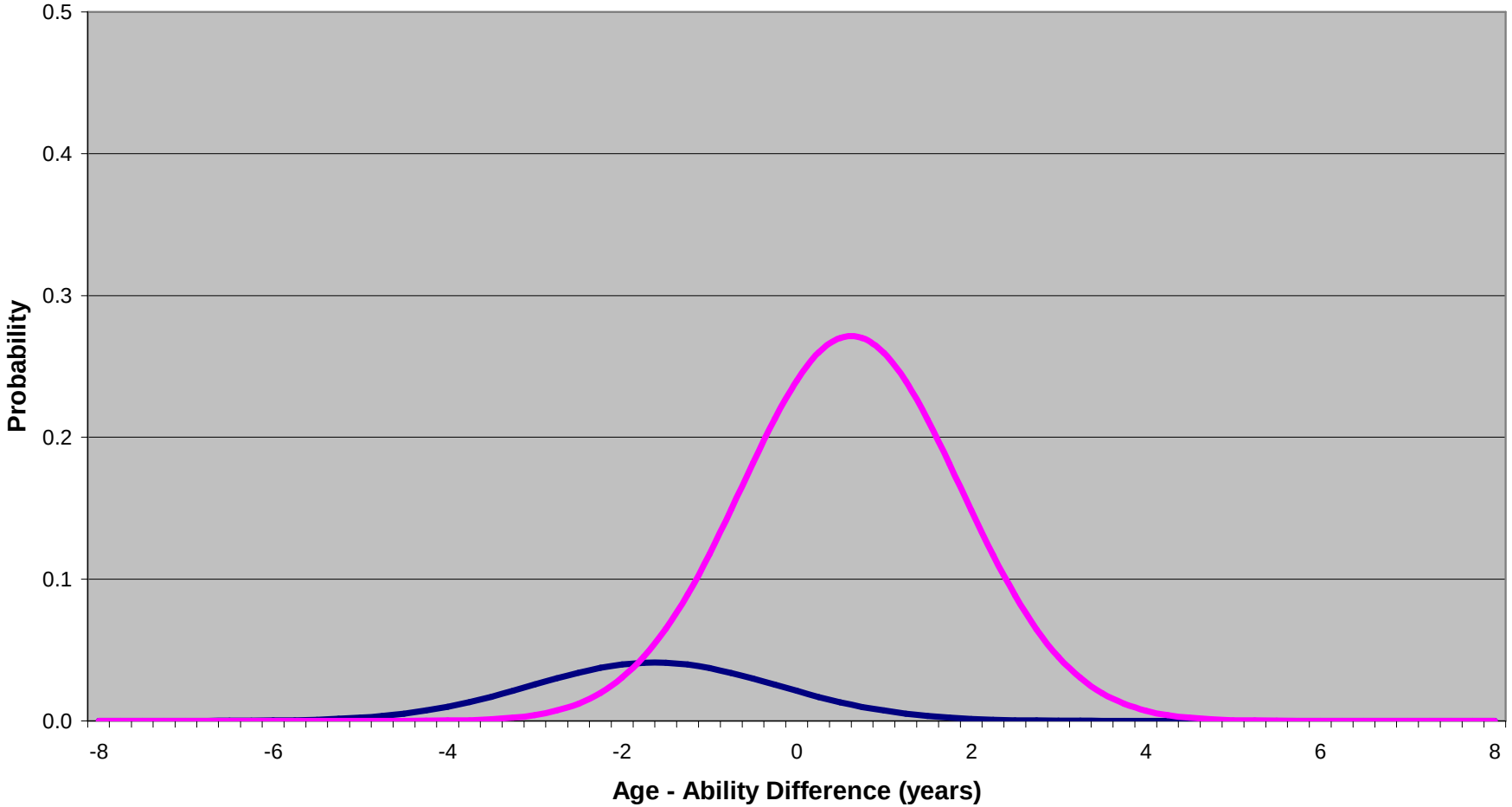


Figure 4.2.3.1: Binormal Subpopulation Plot for Boys' Reading Results in P6

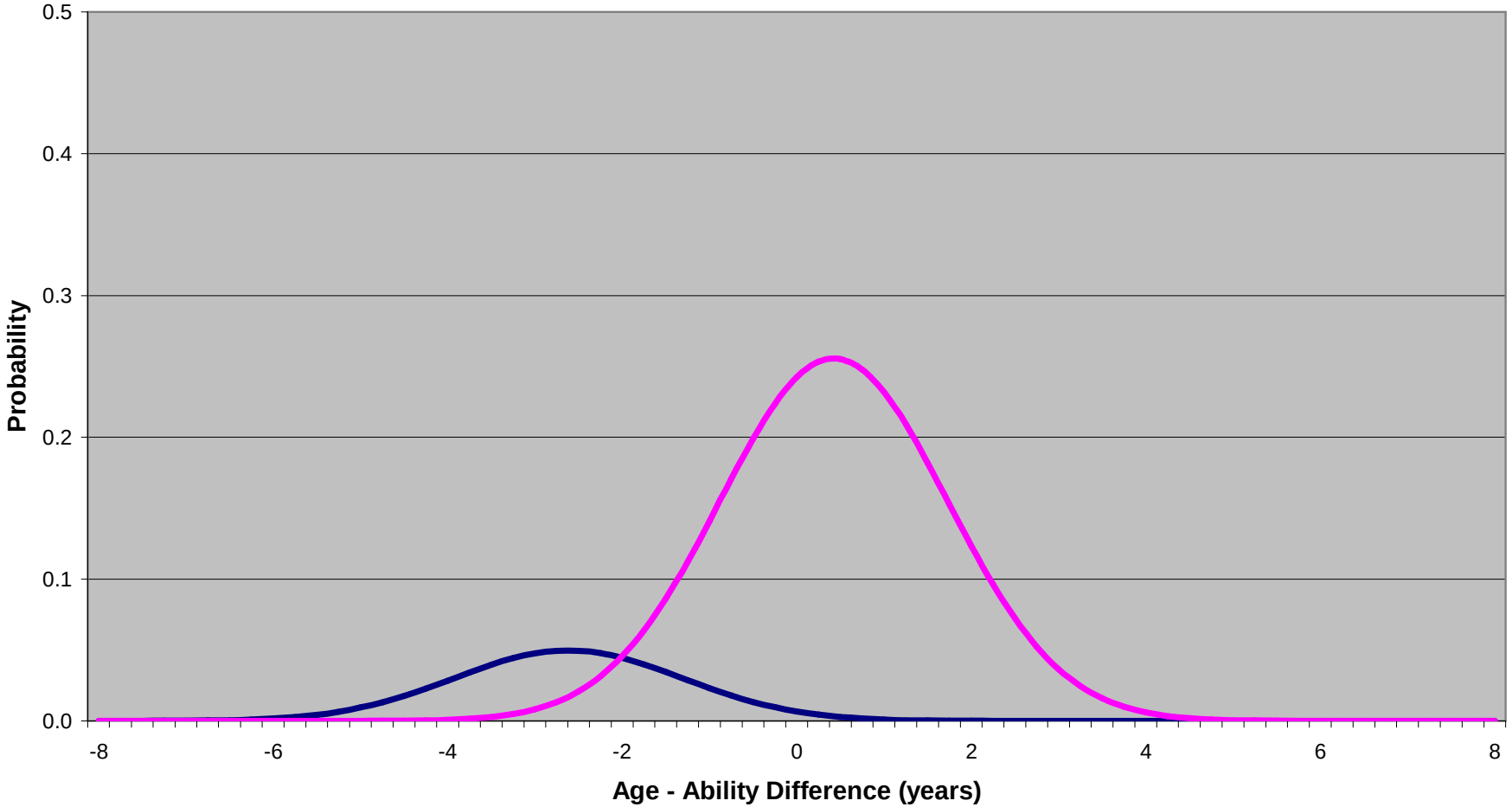


Figure 4.2.3.2: Binormal Subpopulation Plot for Girls' Reading Results in P6

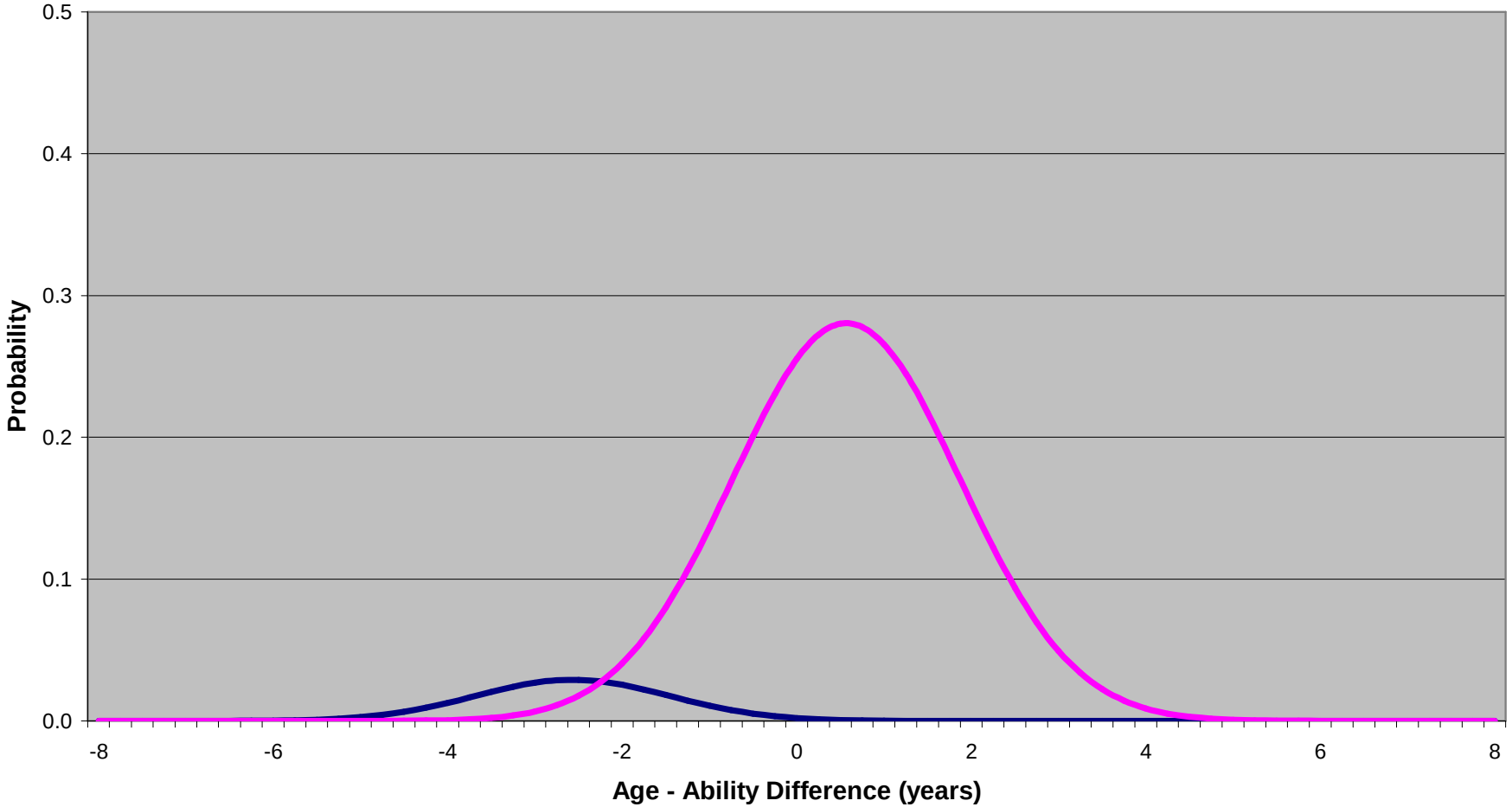


Figure 4.2.4.1: Binormal Subpopulation Plot for Boys' Reading Results in P7

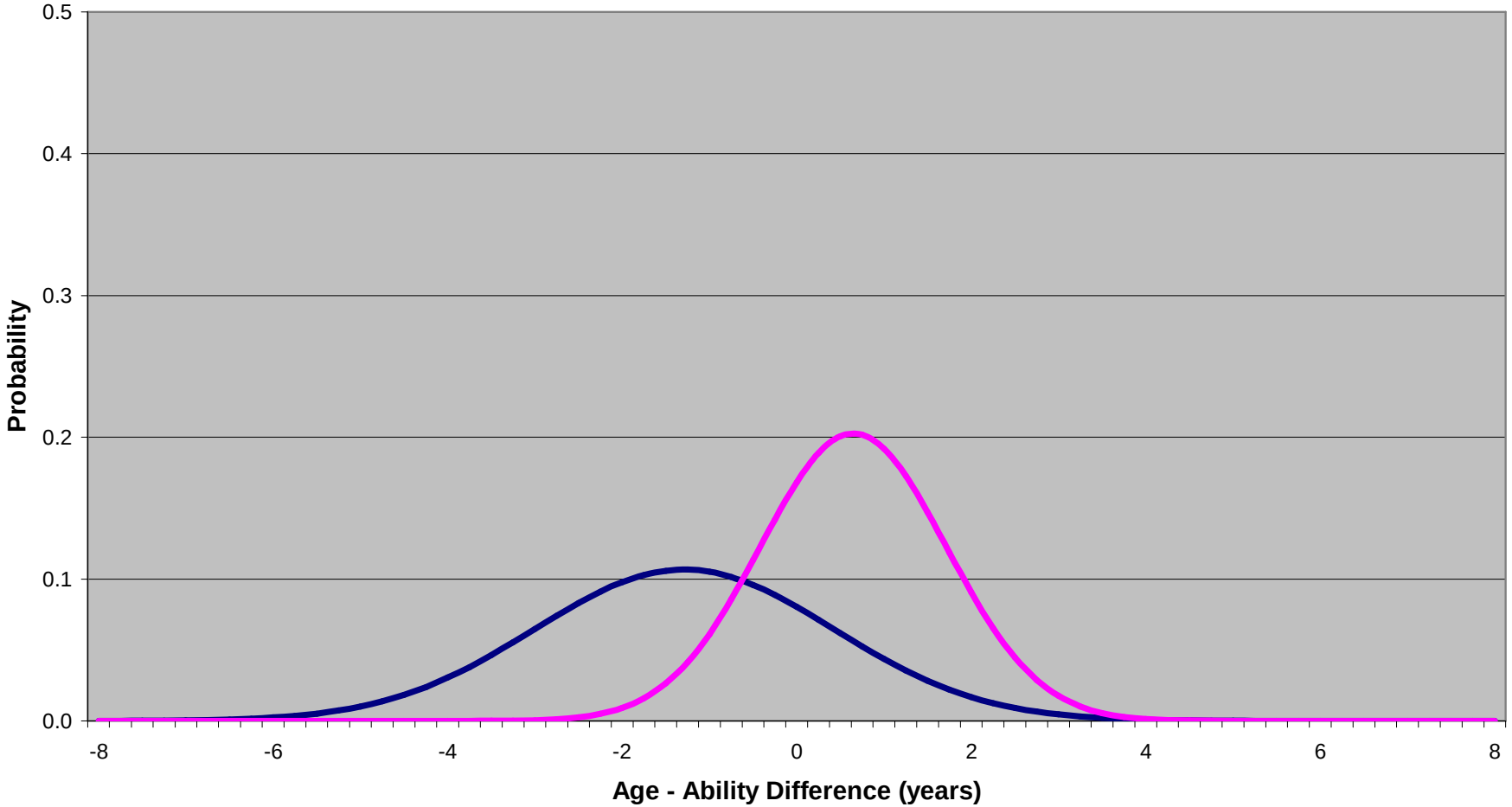


Figure 4.2.4.2: Binormal Subpopulation Plot for Girls' Reading Results in P7

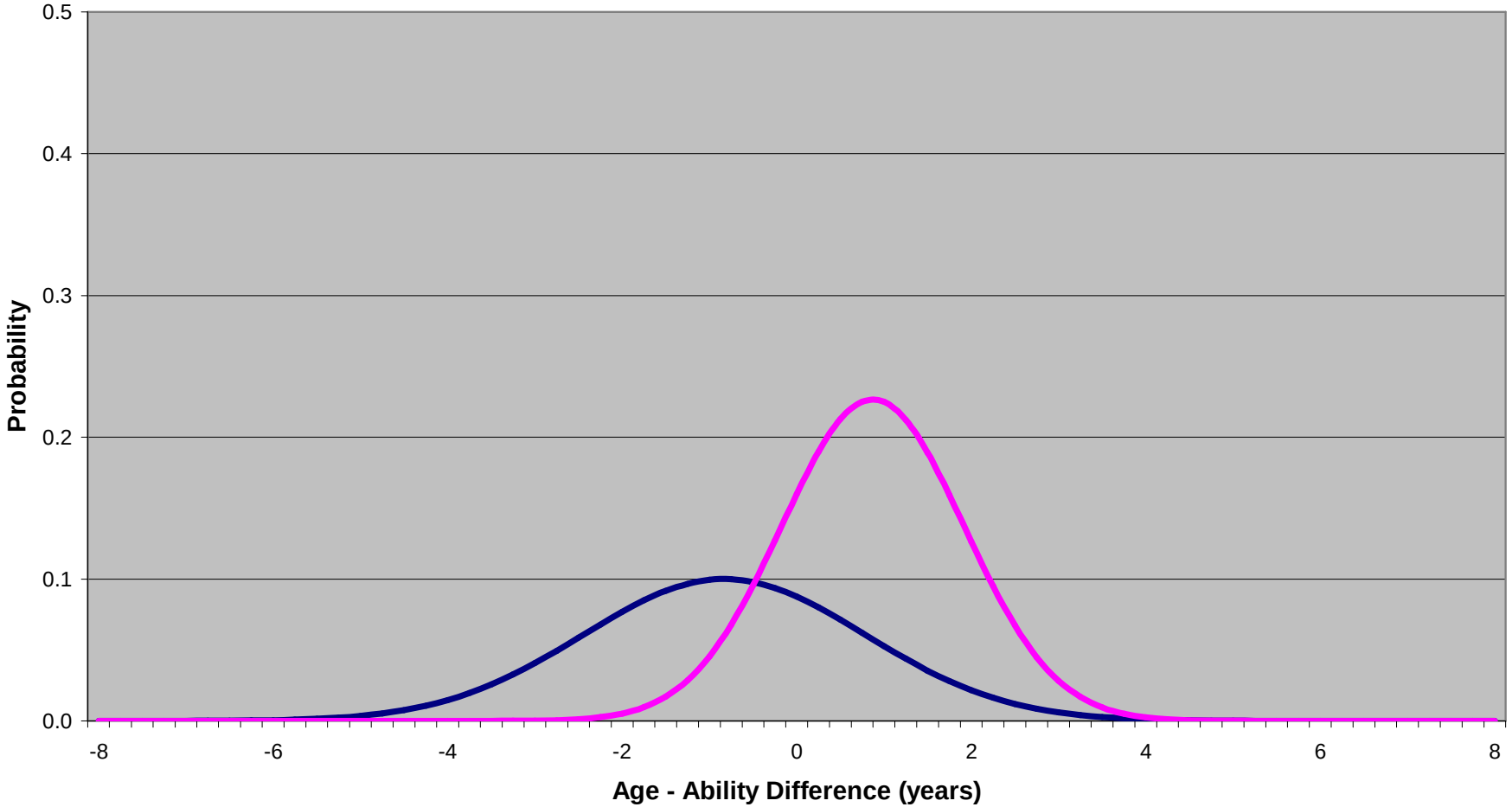


Figure 4.3.1.1: Binormal Subpopulation Plot for Boys' Mathematics Results in P4

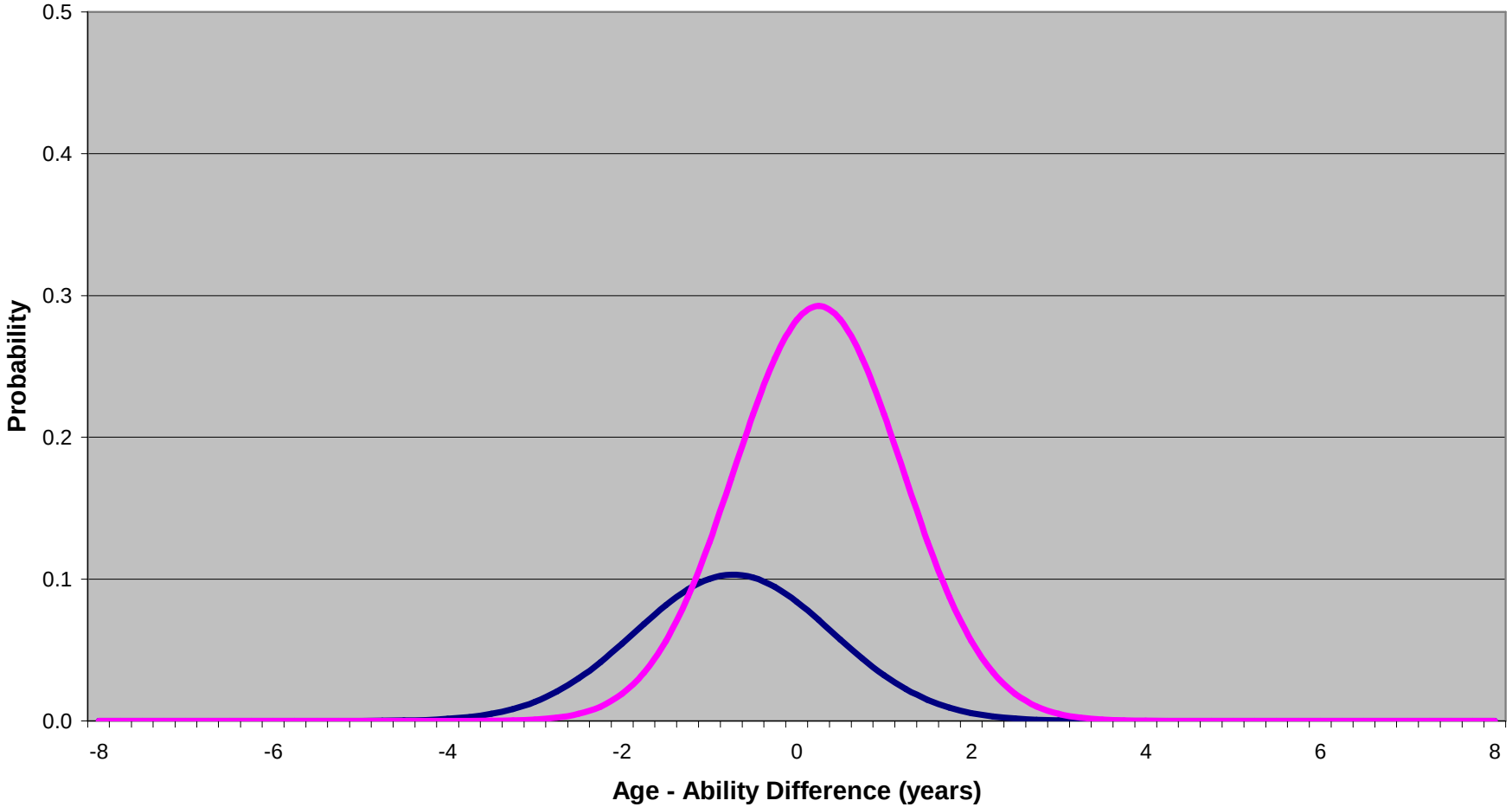


Figure 4.3.1.2: Binormal Subpopulation Plot for Girls' Mathematics Results in P4

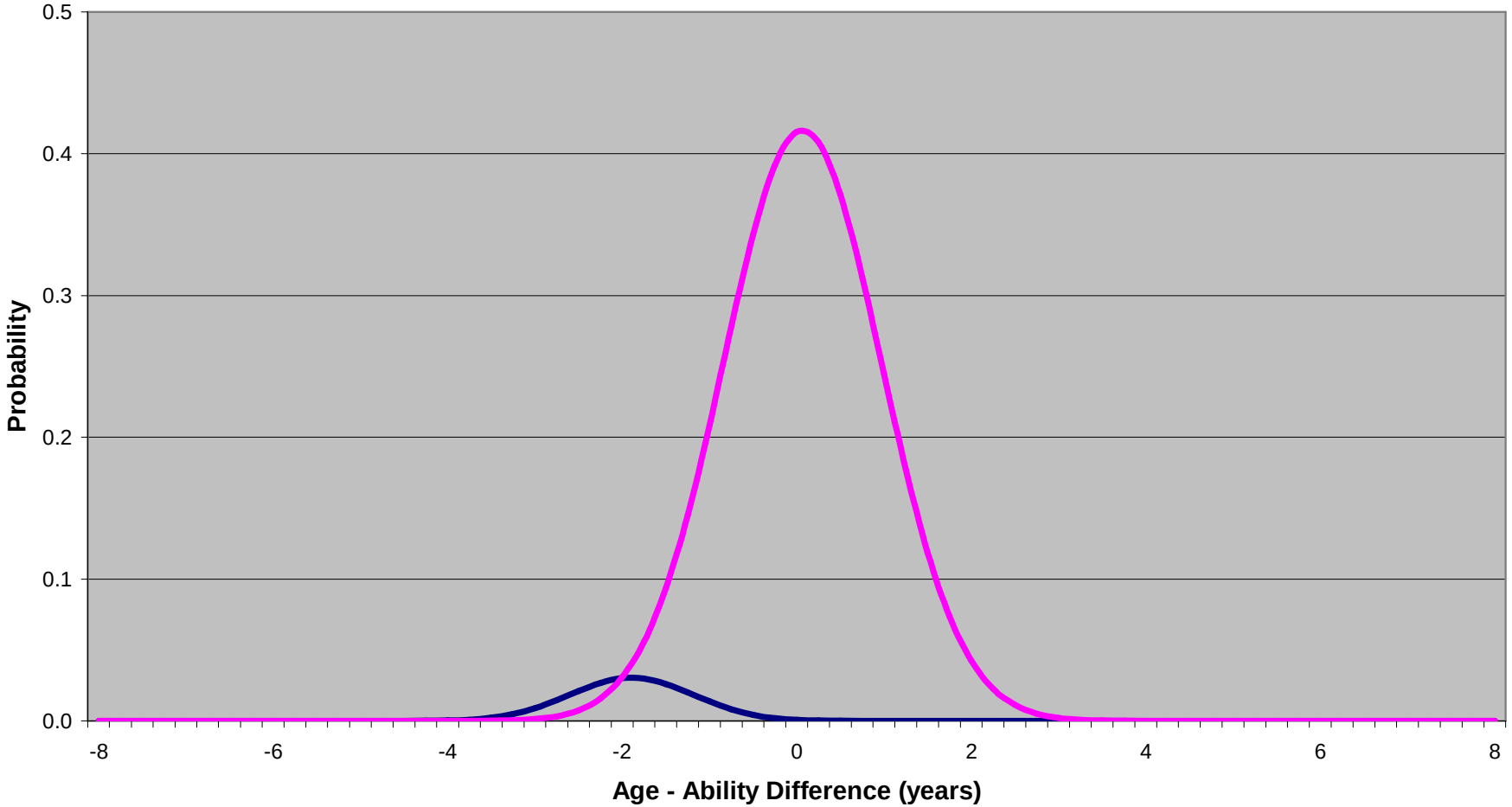


Figure 4.3.2.1: Binormal Subpopulation Plot for Boys' Mathematics Results in P5

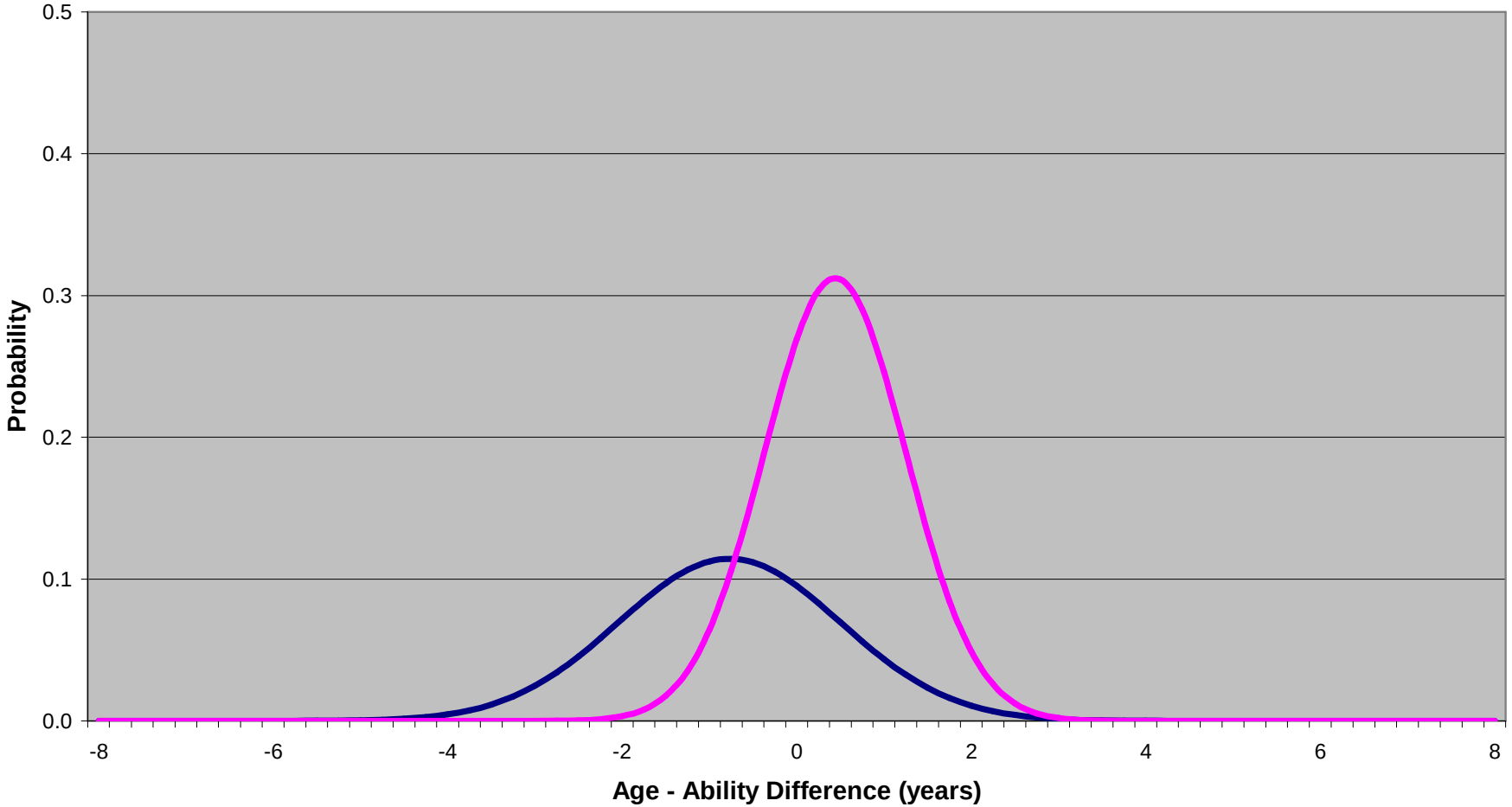


Figure 4.3.2.2: Binormal Subpopulation Plot for Girls' Mathematics Results in P5

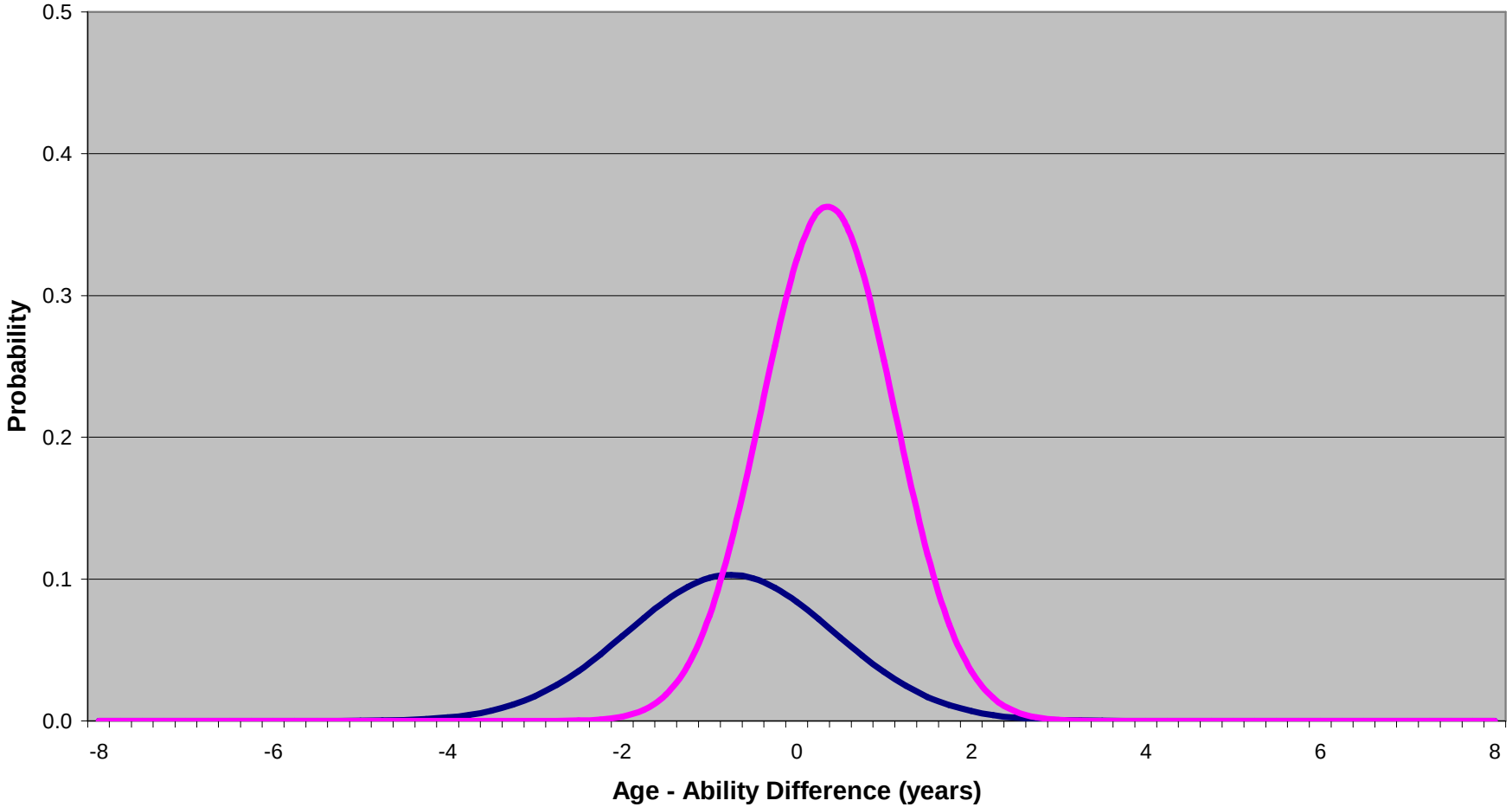


Figure 4.3.3.1: Binormal Subpopulation Plot for Boys' Mathematics Results in P6

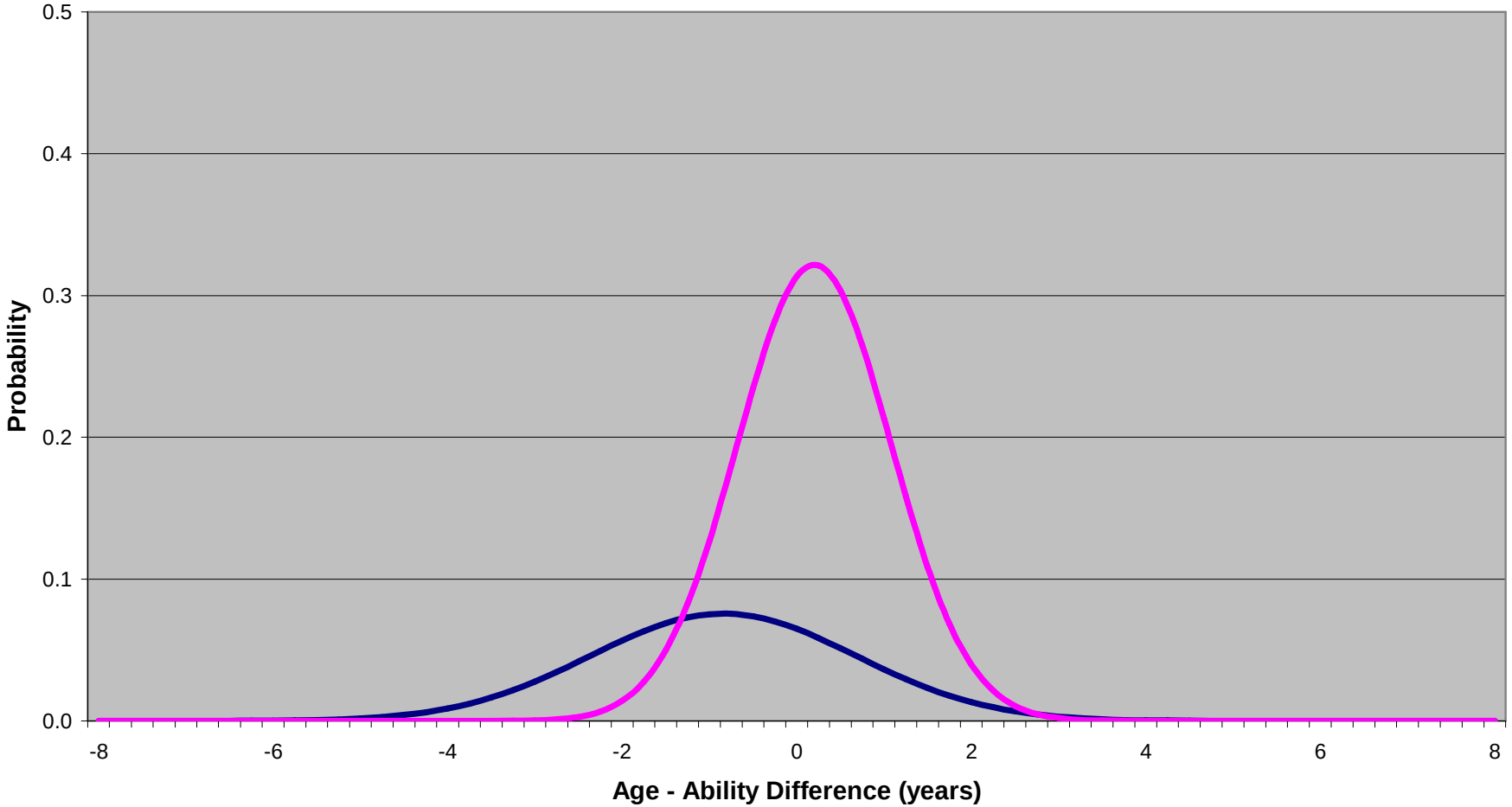


Figure 4.3.3.2: Binormal Subpopulation Plot for Girls' Mathematics Results in P6

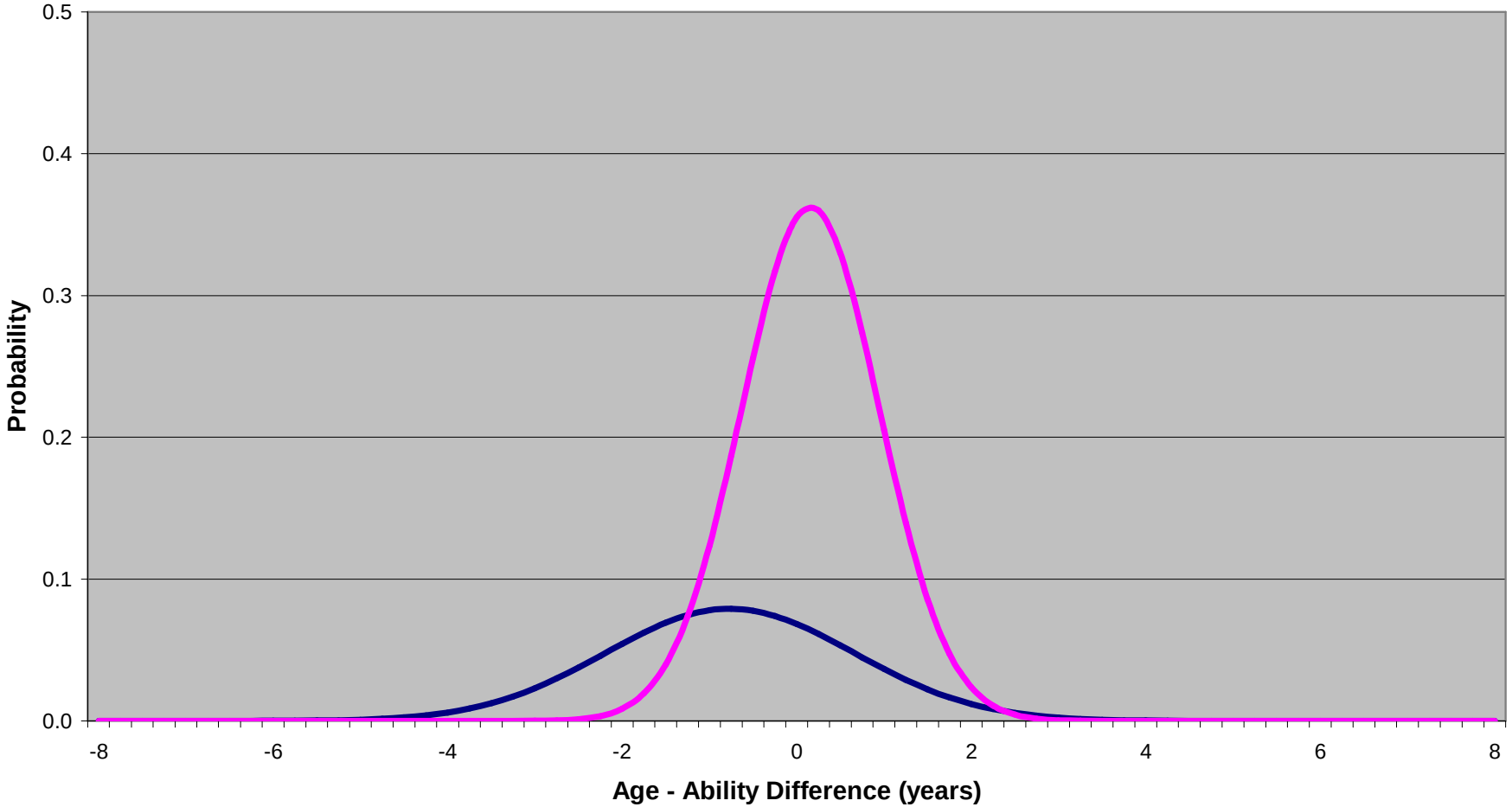


Figure 4.3.4.1: Binormal Subpopulation Plot for Boys' Mathematics Results in P7

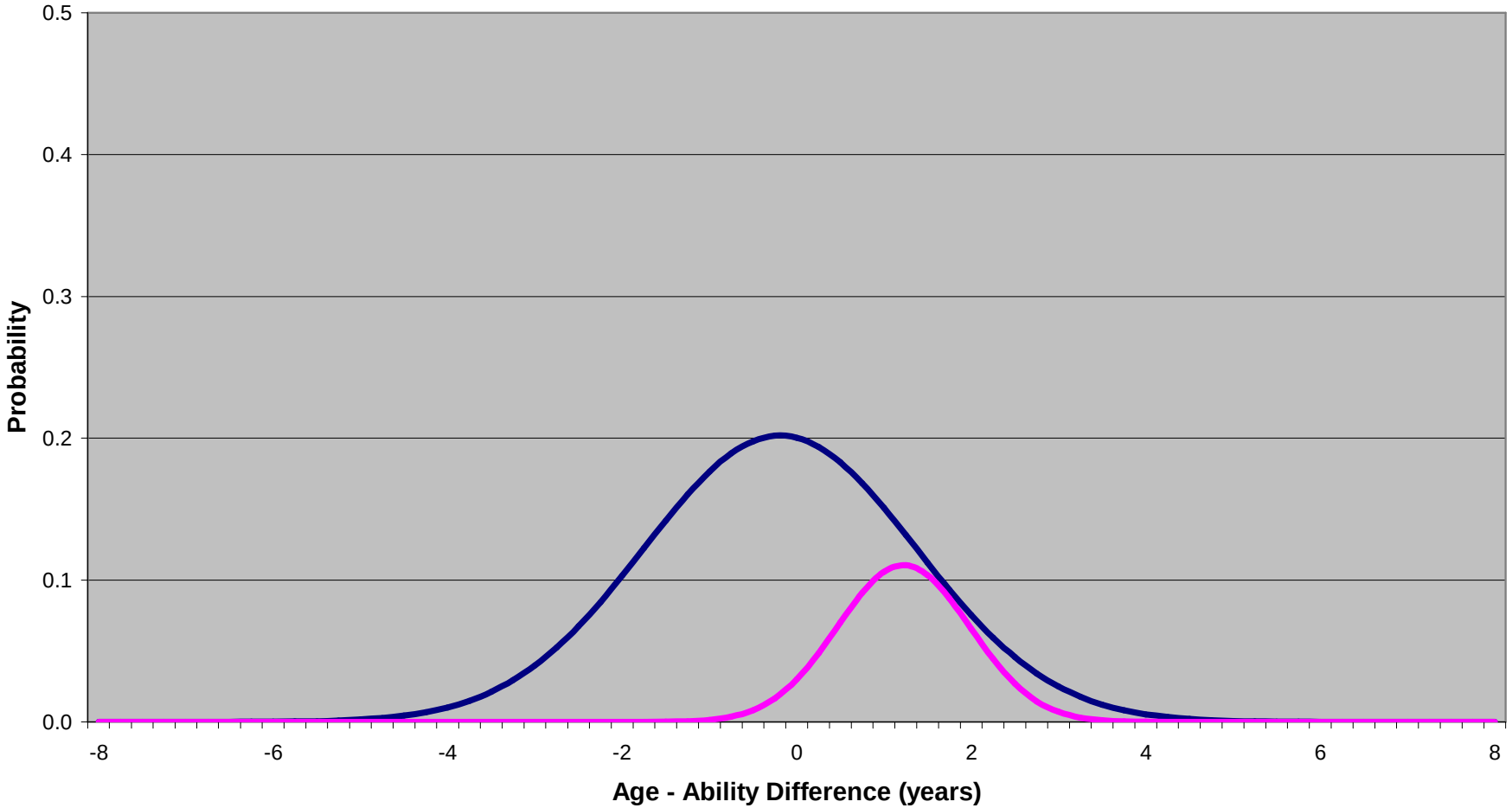


Figure 4.3.4.2: Binormal Subpopulation Plot for Girls' Mathematics Results in P7

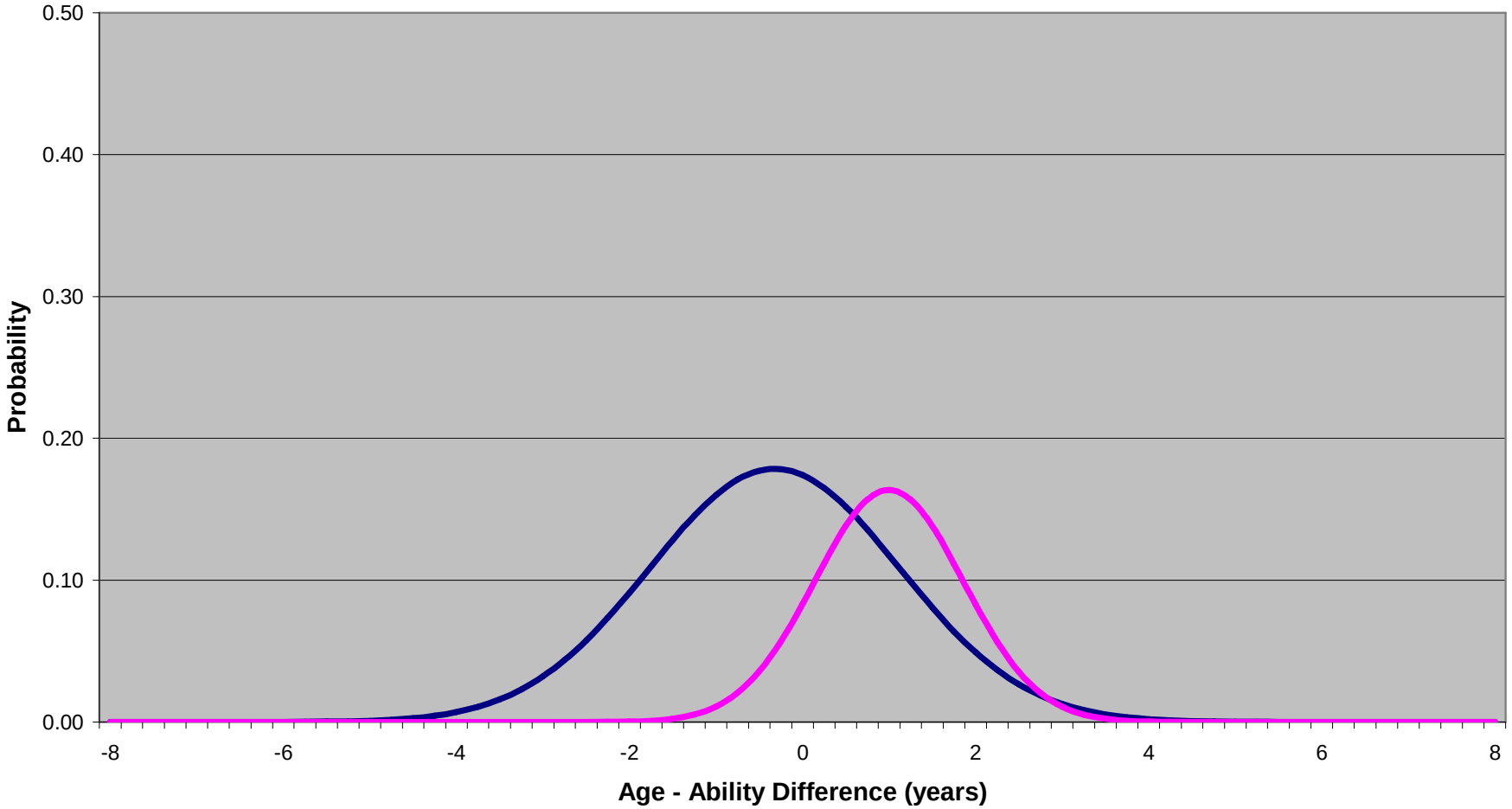


Figure 4.4.1.1: Binormal Subpopulation Plot for Boys' Arithmetic Results in P4

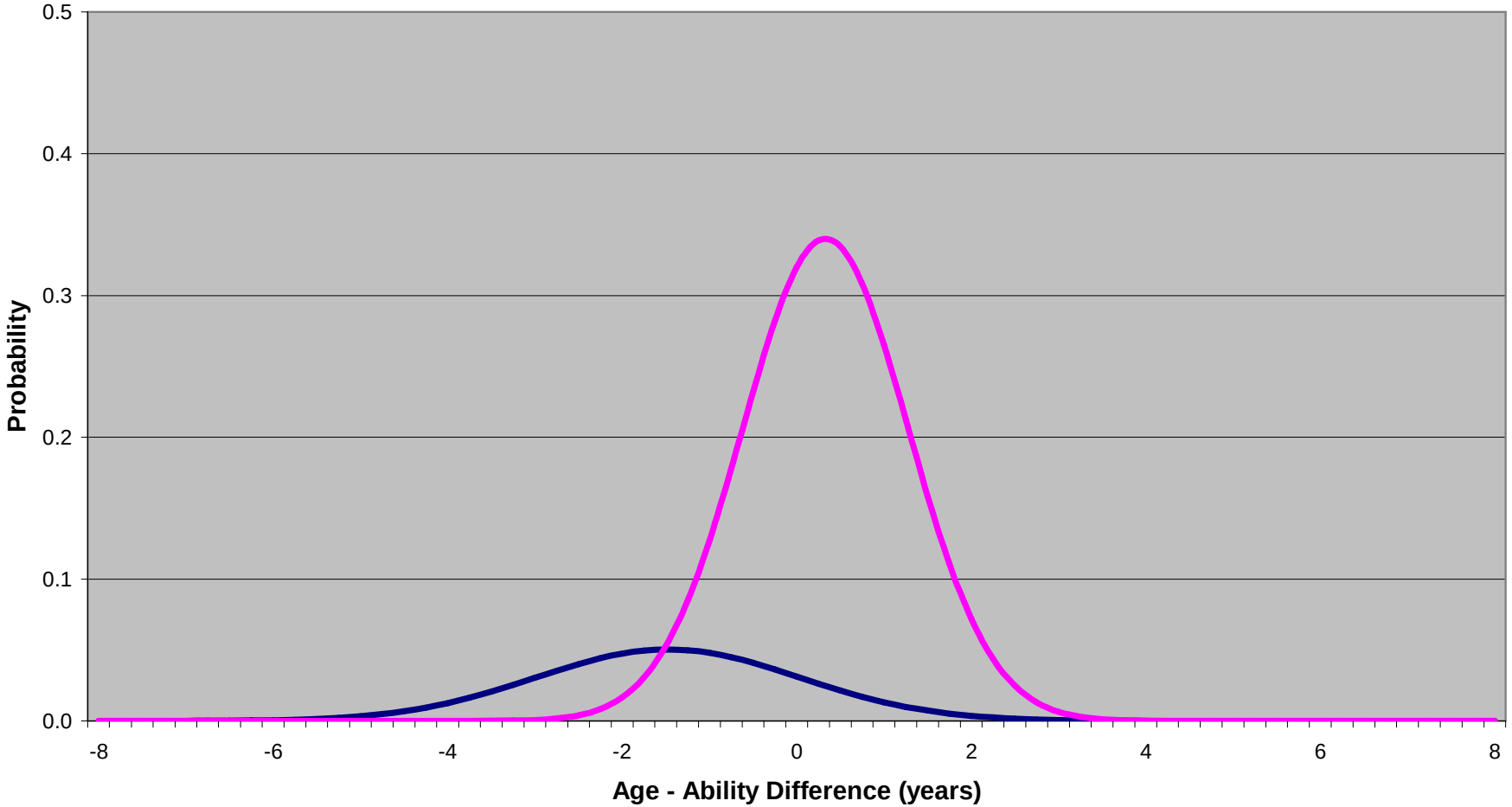


Figure 4.4.1.2: Binormal Subpopulation Plot for Girls' Arithmetic Results in P4

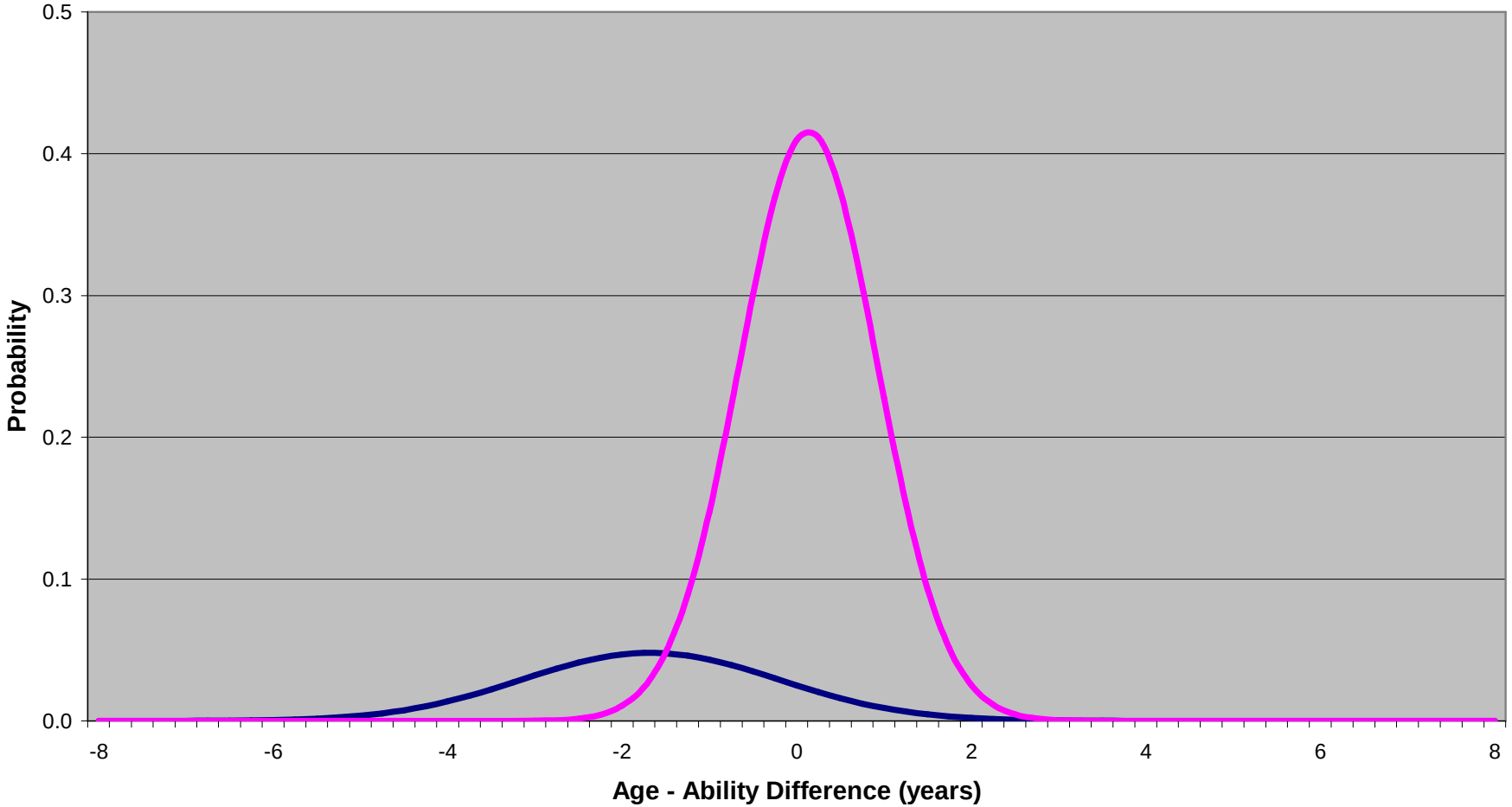


Figure 4.4.2.1: Binormal Subpopulation Plot for Boys' Arithmetic Results in P5

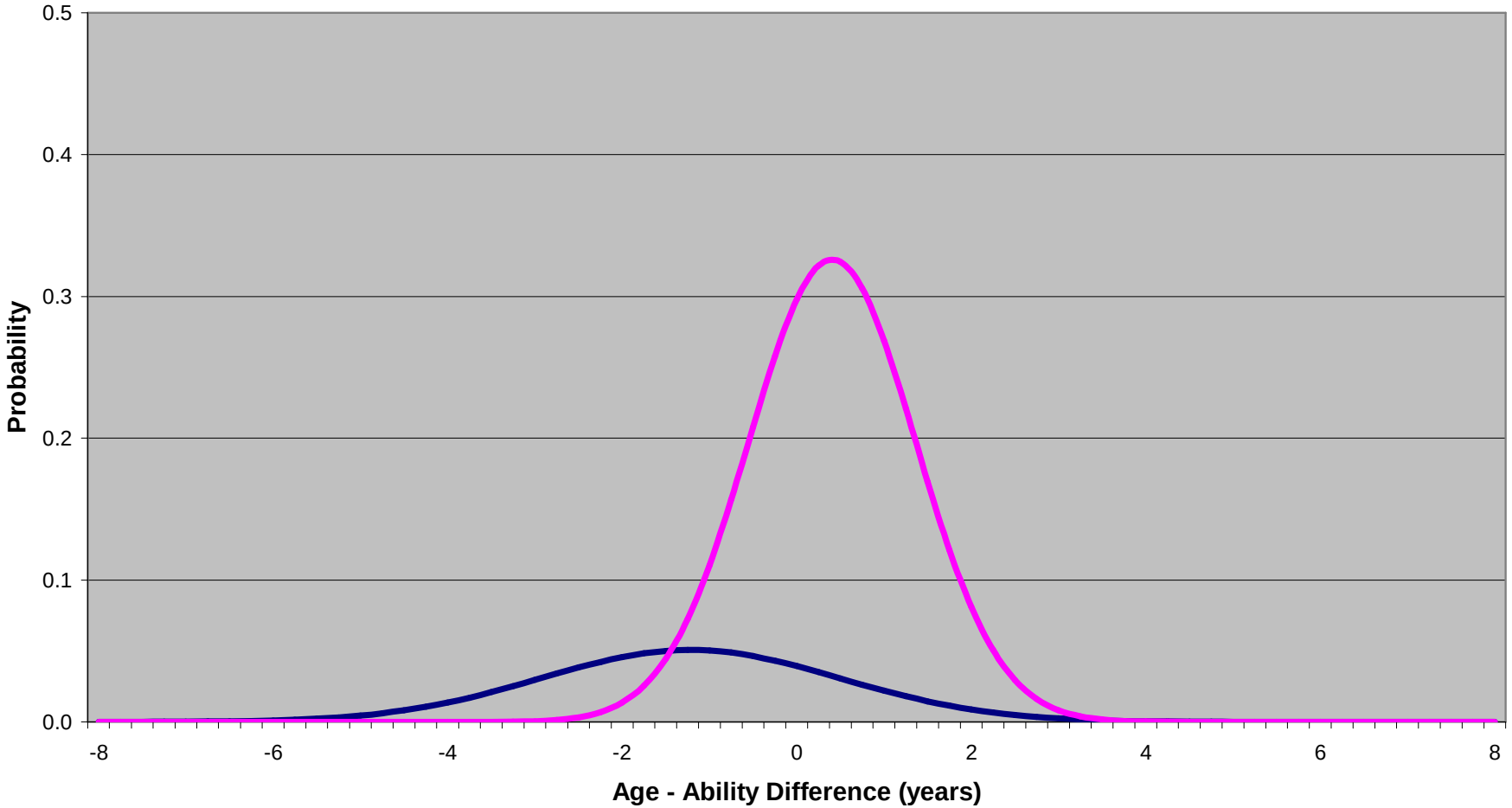


Figure 4.4.2.2: Binormal Subpopulation Plot for Girls' Arithmetic Results in P5

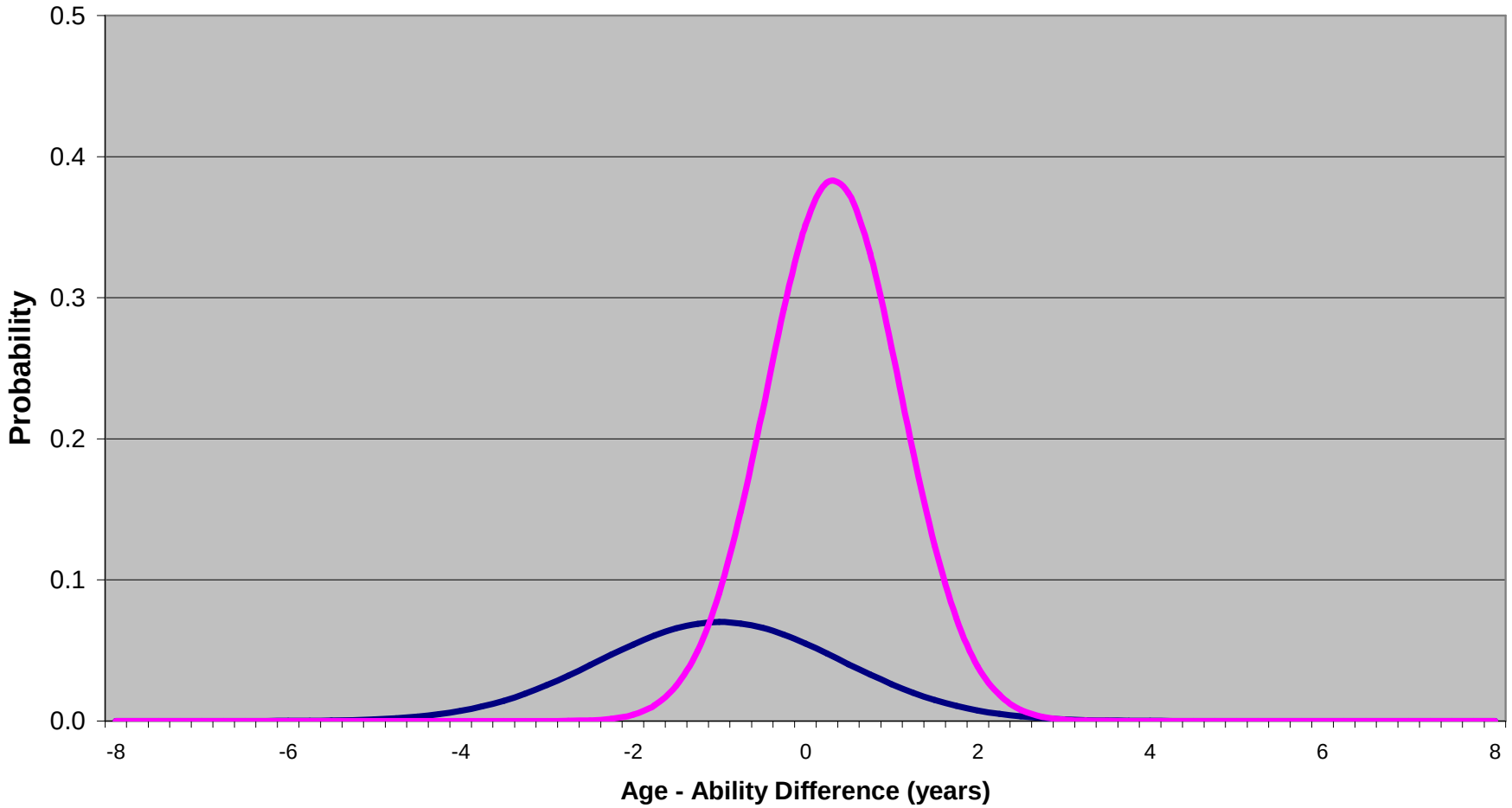


Figure 4.4.3.1: Binormal Subpopulation Plot for Boys' Arithmetic Results in P6

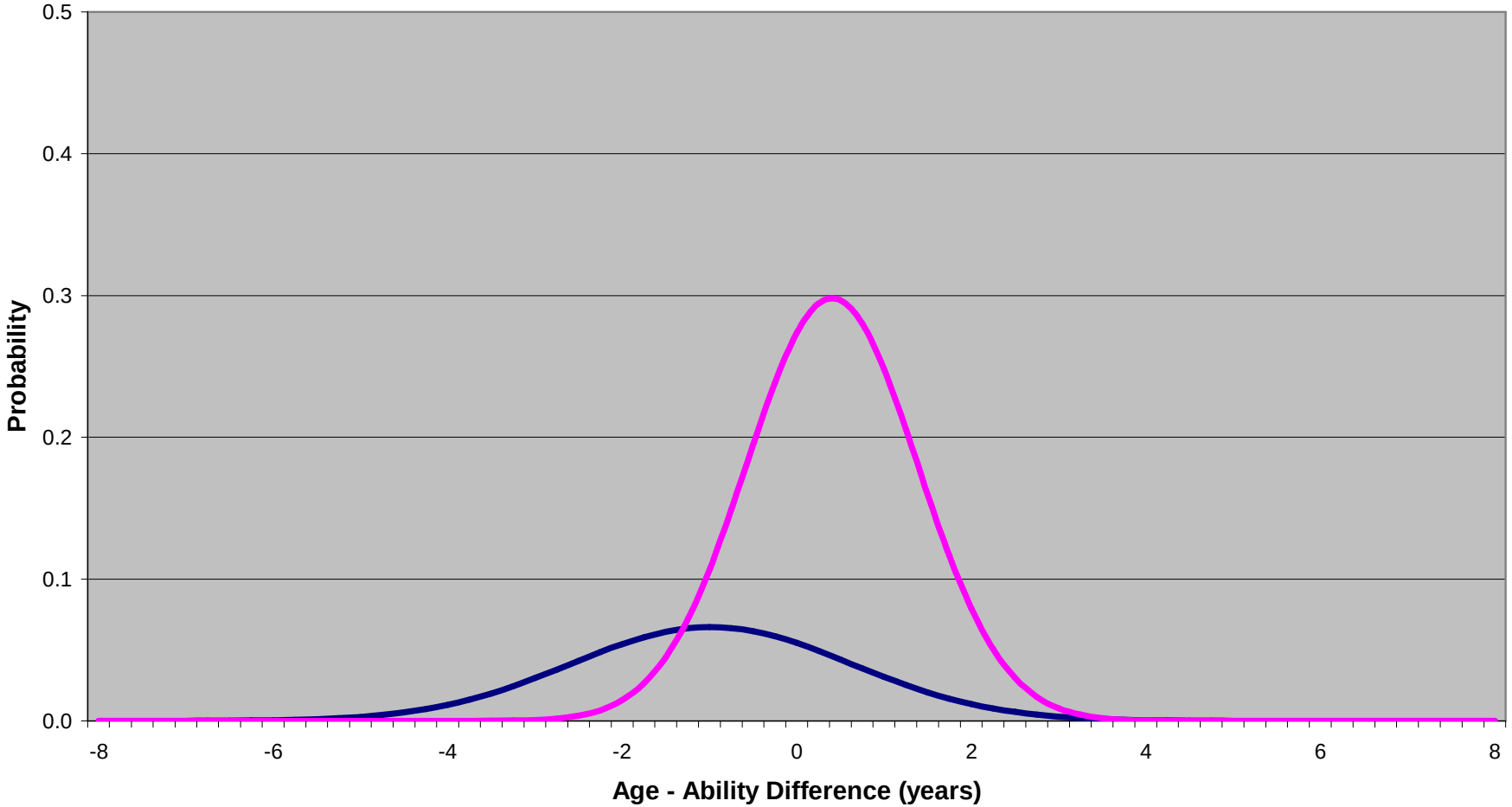


Figure 4.4.3.2: Binormal Subpopulation Plot for Girls' Arithmetic Results in P6

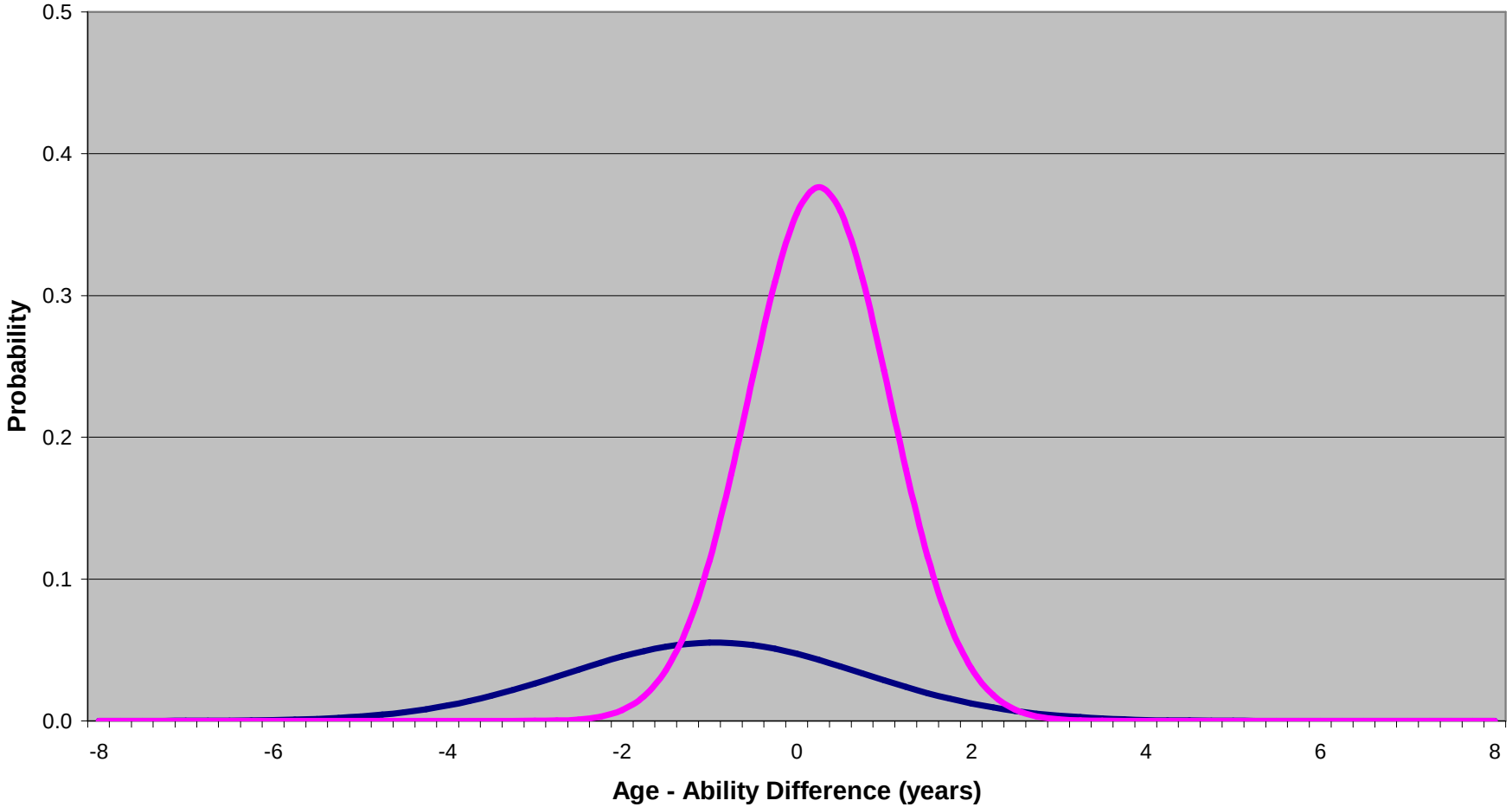


Figure 4.4.4.1: Binormal Subpopulation Plot for Boys' Arithmetic Results in P7

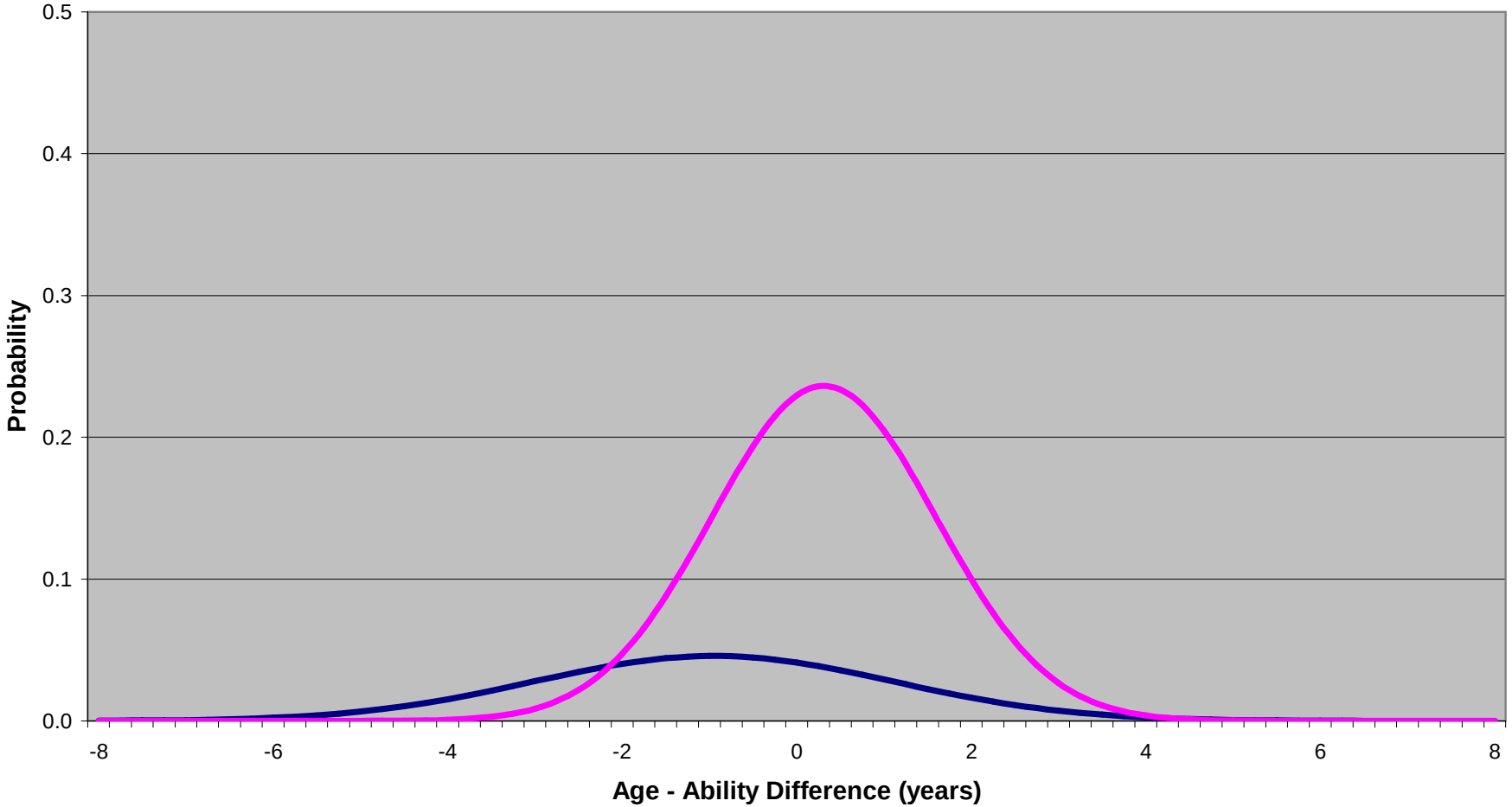
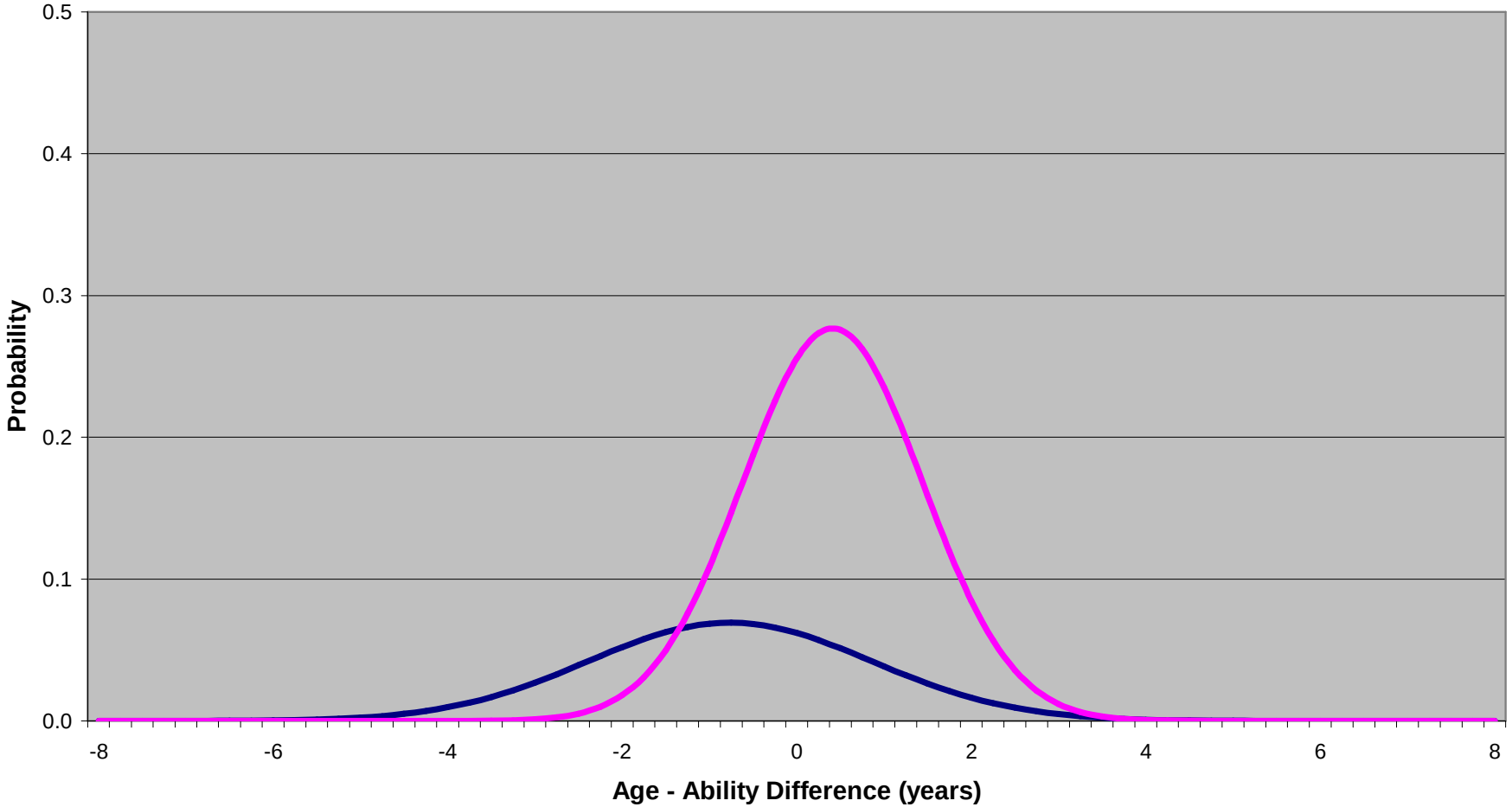


Figure 4.4.4.2: Binormal Subpopulation Plot for Girls' Arithmetic Results in P7



3. Picture Vocabulary

On the whole the model fits illustrated in figure 2 indicate that the normal distribution represents a good description of the observed data. However closer inspection reveals that there is a systematic discrepancy in the fits such that immediately to the left of centre the bars of the histogram tend to be lower than the fitted curve, whilst those immediately to the right of centre tend to be taller. Whilst the normal curve describes a distribution that is symmetrical about the mean, the observed distribution is asymmetrical. It is in fact negatively skewed, having a relatively long and thick left hand tail. This pattern is a consistent feature of the picture vocabulary distributions.

Figure 3 illustrates how well the negative skew in the distribution of scores is described by the binormal distribution. The path of the binormal pdf curve tracks the height of the histogram bars very closely. Figure 4 demonstrates how this improvement in fit is achieved. In all cases the lower attaining subpopulation has a very broad flat distribution which effectively fills the left hand tail, but also extends well in to the right hand side of the distribution. In most cases the standard deviation of the scores of the low attaining subpopulation is actually greater than that of the population as a whole. The one exception to this is in the case of girls' scores in P6, but here the standard error of the estimate is very high and so the exception may be misleading.

Figure 3 also clearly illustrates that with respect to subpopulation prevalence's there is no consistent pattern, either with girls compared to boys or with trends across cohorts.

4. Reading

An examination of figures 2.2.1.1 and 2.2.1.2 reveals little evidence for skewness in the P4 reading data. In each case the height of the histogram falls short of the normal pdf curve towards the centre of the distribution, but a little taller to either side. This negative kurtosis is more clearly illustrated in the boys' data. In the remaining cohorts the data appear to be negatively skewed.

Figure 3 reveals that the binormal model goes some way to describing the negative skew in the data, but that systematic discrepancies between the histograms and pdf curves indicate additional structure in the distributions that are not fully accounted for by the binormal model. This is most clearly evident in the P7 distributions where there is a clear suggestion of a third peak in the distribution of scores.

The subpopulation plots (figures 4.2.1.1 to 4.2.4.2) reveal some consistent patterns in the data. In P4 for both boys and girls there is a considerable overlap in the subpopulation distributions, which accounts for the negative kurtosis and lack of skew in the overall distribution of scores. The standard deviation of the scores of the low attaining subpopulation is always lower than that of the population as a whole, which is in contrast to what was found with picture vocabulary. The prevalence of the low attaining population is always higher for boys than for girls within a cohort. In the case of the boys' data in P4 prevalence of the low attaining subpopulation is greater than that of the higher attaining subpopulation. In all other circumstances the reverse is true. For both boys and girls the prevalence of the low attaining subpopulation decreases between P4 and P5, and between P5 and P6. In P7 there is an increase in prevalence, an observation that may be accounted for if the data were to have a more complex, possibly trimodal structure.

5. Mathematics

The normal model plots for mathematics (figures 2.3.1.1 to 2.3.4.2) illustrate negatively skewed distributions for boys and girls in all year groups. In the three younger cohorts the variation in scores is smaller than was observed for both picture vocabulary and reading, resulting in taller thinner distributions. In P7 the scores are more spread out and as with the P7 reading distribution there is a hint of a third peak hidden in the data.

The binormal model plots (figures 3.3.1.1 to 3.3.4.2) indicate a good model fit for boys and girls in P4, P5 and P6. However the model does not adequately summarise the shape of the more complex distributions observed in P7.

The subpopulation plots for mathematics are illustrated in figures 4.3.1.1 to 4.3.4.2. The plots for boys and girls in the three youngest cohorts account for the negative skew with a broad flat distribution of scores in the low attaining subpopulation. This is similar, though not quite as marked, to the pattern observed for the distribution of picture vocabulary scores. In P7 the pattern is quite different. Here there is considerable overlap between the two subpopulations, and that with the higher attainment represents smaller subpopulation. The pattern of subpopulation parameter values between boys and girls is inconsistent, as is the trend across year groups.

6. Arithmetic

The overall pattern in the distribution of arithmetic scores closely mirrors those found for mathematics. The distributions tend to have a relatively small variance and be negatively skewed. The low attaining subpopulation has a high standard deviation that extends well into the left hand tail of the distribution. However unlike the mathematics distributions this pattern is continued in the P7 results. As with the mathematics data there are some hints toward an underlying trimodal distribution, particularly in P7, although not as obviously so.

With respect to the binormal parameter estimates there is some evidence of a mirroring of those results for boys' and girls', but with inconsistencies in the differences of the absolute magnitudes of those estimates. The most striking cross-cohort pattern occurs in the estimate of the mean score for the low attaining subpopulation. For both boys and girls this increases with the age of the cohort.

7. Summary

In this chapter a visual inspection was made of the normal and binormal model fits to the observed distributions of assessment scores. An examination of the probability histograms revealed a tendency for the data to be negatively skewed. In general the binormal model provided a good description of the data for younger cohorts, but there was evidence for a more complex data structure in some instances. In particular the P7 results for reading and mathematics revealed evidence for an underlying trimodal structure. There was a tendency for the binormal model fits to accommodate the negative skew in the data by utilising a broad flat distribution of scores in the low attaining subpopulation.

In general the results described here concur with the statistical evaluation of model fits presented in chapter 5. However the visual examination of the data gave insights into the distribution of assessment scores that were not immediately apparent in the statistical descriptives. In chapter 7 the validity of the model fits within the context of the binormal hypothesis of specific learning disabilities is considered.

Chapter 7: The Validity of the Binormal Model Fits

1. Introduction

On its own a statistically significant fit does not mean that a particular model has validity. To establish that it necessary to consider how the results sit within a theoretical framework. In this chapter the degree to which these data support the hypothesis that the population consists of qualitatively distinct subpopulations is considered. Other factors that may contribute to the shape of score distributions are also discussed.

2. The Evidence from Age-Grade Curves

An interesting feature of the age-grade curves presented in chapter 4 is the sudden step up in scores that was observed for mathematics and arithmetic between P6 and P7 (figures 1.3 and 1.4). This was accompanied by a sudden increase in the variance of scores (tables 1 .3 and 1.4). The pattern is also illustrated in the probability histograms where there is a suggestion of another peak appearing in the right hand tail of the data. This is perhaps most clearly illustrated in the boys' mathematics results in P7 (figure 3.3.4.1). Here there is a clear spike in the distribution at around about 1 year. Compared to the apparently binormal distribution observed in P6 (figure 3.3.3.1) it is as if a third group of pupils had suddenly pulled away to the right. That conclusion is also supported by the subpopulation plot for the same data where the minority subpopulation has suddenly shifted from the left hand tail to the right (figure 4.3.4.1).

A possible explanation for these observations is that it results from a particular quirk of the Northern Ireland education system where the data were collected. Northern Ireland is one of the few remaining regions of the UK that operates a grammar school entry system. A proportion of children in the province will have been given additional coaching in their final year of primary school for the purpose of sitting grammar school entry exams. This may have had a knock on impact that has affected the InCAS scores, either through a boost in curriculum knowledge or a boost in test-wiseness. It is entirely possible that this would have

a more apparent effect on mathematics and arithmetic than it would on reading, although there is evidence for a similar effect in the P7 reading histograms (figures 3.2.4.1 and 3.2.4.2). If that explanation were correct then a simple way to test it would be to repeat the analysis using InCAS data gathered from somewhere where there are no such tests for secondary school entry. Data gathered from an entire education authority in Scotland or England would be ideal for such a purpose.

Another possible explanation is that the appearance of a subtype of high attaining children in the population is a reflection of a genuine developmental step that occurs at around the age of 10 years. If this were true then it is probable that some children will have made that step already at the age of nine years, and others will still not have made the transition by the time that they are 11. In short there ought to be an observable pattern in the data over time, particularly with respect to prevalence estimates. If such a pattern were found it would be desirable to establish whether it was a specific trend associated with either mathematics or reading, or something more general. It is also entirely feasible that there would be a gender difference in any trend, particularly if it were linked to the onset of puberty. Within the context of the present study finding such a pattern would be dependent upon applying models of increasing modality, starting with the trinormal distribution.

The possibility that the distribution of scores may be affected by instructional factors is not very surprising. Indeed it was acknowledged to be a possible explanation for the skew observed in the Isle of Wight study data (Rutter and Yule, 1975). However this need not be an obstacle to finding evidence for specific learning disabilities. If there are qualitatively distinct subtypes of learner then it is likely that instruction will have a differential effect on those subtypes. Children with specific learning disabilities will still have a tendency to be located at the lower end of the distribution, and that will be apparent in a large enough sample.

3. Picture Vocabulary

In terms of qualitatively distinct subtypes of vocabulary acquisition the major difference might be expected to be found between those that speak English as their first language, and those that speak it as an additional language. The variation in vocabulary scores might be expected to be relatively high for additional language speakers for a variety of reasons. The amount of exposure to English would contribute significantly to this variation. Children that had newly arrived in the country and had only been learning English for a short while would be at a considerable disadvantage compared to those born in the country that, whilst having a different mother tongue, had had some exposure to English for the whole of their lives. In addition picture vocabulary assessments are known to have a cultural load that is likely to disproportionately affect recent immigrants. As well as a relatively high variation in scores it is also likely that there would be considerable overlap between the two subpopulations. Whilst it might be expected that the vocabulary score of the weakest native English speaker might not be as low as that of the weakest additional language speaker, there is no reason to suppose that there should be the same difference at the other end of the scale.

At first glance the picture vocabulary data presented here are largely consistent with the expectation presented in the previous paragraph. However closer examination of the data reveals a problem. If new immigrants continue to arrive in the country it will tend to hold down the lowest vocabulary score of the additional language subpopulation, while the highest score for the same subpopulation will tend to increase. This would result in a larger variance in the scores of the additional language subpopulation as the cohorts increase in age. This pattern is not found in the data (tables 9.1.1 and 9.1.2). It may be that additional language speakers do not form a high enough proportion of the population to make a noticeable impact on the distribution of scores. Although reliable figures for the proportion of EAL (English as an Additional Language) children in the sample were not available, it is reasonable to suppose that Northern Ireland represents a region of the UK with a relatively low number of such children.

Year Group	Variable				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\bar{D}}$	$s_{\bar{D}}$
P4	0.07	-3.27	2.50	0.07	1.68
P5	0.53	-0.37	2.16	0.50	1.27
P6	0.23	-1.12	2.43	0.53	1.55
P7	0.28	-0.97	2.55	0.55	1.67

Table 9.1.1: Summary of binormal parameter estimates for boys' picture vocabulary scores

Year Group	Variable				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\bar{D}}$	$s_{\bar{D}}$
P4	0.27	-1.43	1.91	0.42	1.37
P5	0.30	-0.70	2.42	0.35	1.26
P6	0.03	-5.00	1.70	0.12	1.61
P7	0.18	-1.80	2.16	0.22	1.74

Table 9.1.2: Summary of binormal parameter estimates for girls' picture vocabulary scores

If EAL children are not making a significant contribution to the distribution of picture vocabulary scores then it might be expected that they would follow a normal distribution. In fact this doesn't appear to be the case if the additional variation explained by the binormal model fit is taken into consideration (table 8). This shows that fitting the binormal model to the picture vocabulary data has a greater impact than it does on either that of reading or mathematics. However this might simply have resulted from having four times less data which will have affected the smoothness of cdf curve used in the model fit. When the parameter fit statistics were considered the binormal distribution model was rejected in favour of the normal distribution model for girls' picture vocabulary scores in P7 (table 7.1.4.2). However, this aside the binormal distribution model did provide a significantly better fit. If the negative skew in the data cannot be explained in terms of the presence of EAL children then it suggests that something more complex is happening.

Early vocabulary acquisition is achieved entirely by listening to language. However as a child learns to read they increasingly acquire vocabulary through the printed word. The degree to which children acquire vocabulary from print will depend upon their reading ability and their propensity to read. Thus children with specific reading disabilities are likely to acquire less vocabulary through

print, but that is not to say that they do not develop compensatory mechanisms for language development. The vocabulary development of normal readers that never pick up a book will also be affected. Thus the distribution of picture vocabulary scores may be intimately connected with reading development in complex ways.

If vocabulary acquisition is indeed linked to reading development then the connection will be most prominent in data collected from older children. A tantalising glimpse of that connection is revealed if the P7 girls' probability histograms for picture vocabulary and reading are compared (figures 3.1.4.2 and 3.2.4.2). In this case the binormal distribution model for picture vocabulary was rejected on the grounds that mean and prevalence estimates of the low attaining population were not significantly different from zero. However the distribution of scores is clearly not unimodal, and perhaps a trinormal distribution would provide a better fit. There is also a suggestion that the distribution of reading scores might be better described by a trinormal distribution.

When cross-gender and cross-cohort comparisons are made of the binormal parameter estimates then no particular patterns are observed. This may result from complex interactions with reading development and EAL status. If so this exposes an important limitation of the methodology to cope with such complexity.

4. Reading

The distributions of reading results appear to have a complex structure that is not adequately explained by the binormal distribution model. It has been argued that in P7 an apparent boost in mathematics and arithmetic scores may have resulted from the preparation of some learners to take grammar school entrance tests, and that has introduced additional complexity into the distribution of assessment scores. However this explanation would seem to be inadequate when applied to reading. On the one hand there was little evidence for a boost in reading scores in P7. It might also be expected that the development of reading skills requires

practice over extended periods and may therefore be less susceptible than mathematics and arithmetic to short term cramming techniques.

A second observation is that the apparently complex structure in the distribution of reading scores is already apparent in younger cohorts of children. For example the distribution of girls' reading scores in P5 appears to show three separate peaks at around -2, 0 and 1 years (figure 3.2.2.2). A possible explanation for this distribution of scores is that it results from the relatively complex structure of the reading assessment, composed as it is of separate subtests of word recognition, word decoding and reading comprehension. The apparent trimodality in the results might simply be an artefact. Clearly one way to check this would be to analyse the results for each subtest separately.

Another explanation for the complexity of the distribution of reading scores is that it is a true reflection of the distribution in the population, and that the composite nature of the reading assessment used here has simply revealed it. The simple view of reading (SVR) is a widely investigated model of reading disabilities originally proposed by Gough and Tunmer (Gough and Tunmer, 1986). According to this model reading comprehension is the product of two quite different skills; word decoding and listening comprehension. The model predicts two types of specific reading disability according to which of these skills are impaired. It is suggested that a deficit in word decoding results in the condition usually referred to as dyslexia, whilst a deficit in listening comprehension manifests itself as hyperlexia. There is a growing body of evidence to support the independence of word decoding and listening comprehension in the development of reading, thus providing support for the SVR model (Kendeou et al., 2009). The InCAS reading assessment employed in the present study does not include an assessment of listening comprehension, although it does include word decoding. If the SVR model is correct then a trimodal distribution of InCAS reading scores might be expected. The lowest attaining subpopulation would be those children that were unable to decode (dyslexics). Children that could decode but not comprehend (hyperlexics) would appear as a hump in the middle of the distribution. Normal readers would appear as the highest attaining subpopulation. It is likely that in younger cohorts the

presence of dyslexics might be evident, but that hyperlexics would be indistinguishable from weaker normal readers. However as the cohorts increased in age and reading development it is likely that the hyperlexic and normal readers would gradually separate out, resulting in an increasingly trimodal distribution of scores.

If the data do indeed have a trimodal structure, or perhaps one of even higher modality, it presents a fundamental difficulty when it comes to fitting the binormal distribution model. Unable to cope with the complexity in the data the model is likely to find a best fit solution that merges particular sub-distributions within the whole, resulting in misleading parameter estimates. That means that there will always be an element of doubt into to validity of those estimates. However one way in which our confidence in the parameter estimates can be enhanced is if patterns are found across different datasets. When this kind of reasoning is applied to the reading data it reveals some interesting results. In the context of the present study it is possible to look for differences and similarities between boys and girls, or to look for trends across cohorts.

Tables 9.2.1 and 9.2.2 summarise the parameter estimates for reading that were originally presented in table 7.2. The binormal parameter estimates for the girls' reading scores are considered first of all. There appears to be a clear trend in the results from P4 to P6. The prevalence of the low attaining subpopulation is around 37% in P4 and roughly halves each year to 8% in P6. At the same time the mean score of the same subpopulation decreases by about 1 year each year. If the mean score is added to the mean age at test (table 1.2) it is possible to calculate the mean reading age of the low attaining subpopulation in each cohort. The figures come out as follows; 6.96 years in P4, 7.14 years in P5, and 7.21 years in P6. It is as if the low attaining group are defined by a reading age of about seven years. These observations are consistent with the notion that there is a fundamental developmental transition at about the age of seven. What that may be cannot be stated on the basis of this evidence, but one possibility is that it represents the transition from a beginning reader to a fluent one. Further clues to test that hypothesis might be found in an analysis of the reading subscale results.

Year Group	Variable				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\bar{D}}$	$s_{\bar{D}}$
P4	0.76	-0.87	1.47	1.18	1.18
P5	0.17	-2.53	1.25	0.32	1.37
P6	0.16	-2.62	1.31	0.42	1.31
P7	0.46	-1.29	1.71	0.65	1.07

Table 9.2.1: Summary of binormal parameter estimates for boys' reading scores

Year Group	Variable				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\bar{D}}$	$s_{\bar{D}}$
P4	0.37	-0.82	1.37	0.61	1.27
P5	0.15	-1.63	1.41	0.62	1.26
P6	0.08	-2.57	1.12	0.57	1.31
P7	0.41	-0.83	1.62	0.87	1.05

Table 9.2.2: Summary of binormal parameter estimates for girls' reading scores

Whatever the validity of the speculative hypothesis presented in the previous paragraph may be, the pattern in the reading results do support the premise that there is a qualitatively distinct group of low attaining readers. If the population contains more than two qualitatively distinct groups, three in the case of a trimodal distribution, then it follows that the parameter estimates given for the higher attaining subpopulation actually represent a summary of the remaining scores, rather than anything meaningful in themselves. In the case of a trimodal distribution these would represent a summary of the middle attaining and high attaining subpopulations. Such reasoning can be used to explain the apparent inconsistency in the P7 results. If the observed trend in which the prevalence of the low attaining group tends to halve each year were to continue into P7, then we would expect to see a prevalence estimate of about 4%. Since the binormal model is unable to resolve three subpopulations, the model fit may have simply have coped by merging what is now a very small low attaining subpopulation with the middle attaining group, and reporting the high attaining group as a distinct subpopulation. Evidence that is the case is found in the estimate of the standard deviation of the higher attaining subpopulation which falls from about 1.3 in the younger cohort to a little over one in P7.

Interpretation of the binormal parameter estimates obtained from the boys' scores is more problematical. At first glance there is a striking consistency between the parameter estimates obtained in P5 and P6. In both cohorts the prevalence figure for the low attaining subpopulation is about 16% and the mean score is about -2.5 years. However if the parameter estimates are compared with those obtained for the girls then there is a certain consistency in the results for the P4, P6 and P7 cohorts, and that would suggest that the P5 results may be misleading. Specifically the means of the low attaining subpopulation are very similar between boys and girls in P4 and P6, whilst the prevalence figure for the boys is about twice that reported for the girls in each case. In P7 the prevalence figure is about 5% higher for boys, but that is consistent with the notion that the prevalence of the low attaining subpopulation halves every year. If the contribution of the low attaining subpopulation to the prevalence figure reported in P7 is 4% for the girls, then by the same logic it will be 8% for the boys, thus accounting for the bulk of the 5% difference. On balance then the latter explanation accounts for more of the observations in the boys' results whilst maintaining consistency with the patterns found in the girls'. Inconsistent observations may be explained by limitations in the methodology for which further investigation is required.

To what extent then do these results support the medical model of specific reading disabilities? Certain cross-cohort and cross-gender patterns in the results provide strong evidence for two qualitatively different subpopulations of reader that are consistent with a developmental transition in reading ability at a mean age of about seven years. The medical model of learning disabilities predicts qualitatively different subpopulations as a result of specific cognitive deficits. In recent years a growing body of evidence has accumulated that reading disabilities may result from such a deficit in phonological processing and that a second subtype of reading disability may exist that is linked to a core deficit in naming speed (Vukovic and Siegel, 2006). It has also been suggested that it is a deficit in visuo-spatial processing that is responsible (Stein et al., 2000, Stein and Walsh, 1997, Stein, 2001, Vidyasagar and Pammer, 2010). Either way if reading disabilities are indeed caused by an innate cognitive deficit that would predict that the prevalence of the condition will be constant across different age groups.

Prevalence estimates vary according the details of the procedures used to identify individuals; however a figure of between 5.4% and 7% has been cited (Snowling, 2005). In the present study the lowest prevalence estimate was found to be higher than this at 8% in the case of 10-year-old girls. Furthermore the prevalence is not constant across cohorts but tends to decrease by about 50% each year. This result suggests that the cause of reading disabilities is a delay in a developmental step that occurs on average at seven-years-of-age. However that does not exclude the possibility that there are children for whom the cognitive systems required for reading are so compromised that they never make that developmental step. Such children might be classed as 'true' dyslexics. This explanation is consistent with the causal model of dyslexia proposed by Morton and Frith (Morton and Frith, 1995). According to this model dyslexia results from the failure at the cognitive level of a critical neurological structure, referred to as 'P'. Failure of this structure may have several causes at the biological level. These include a 'faulty brain system' as predicted by the medical model, or in the case of younger children an 'immature brain system' indicating a developmental cause. If this model is correct then behavioural level observations of reading ability will fail to identify the underlying biological cause of reading failure. Nevertheless it is theoretically possible to extrapolate the trend in prevalence estimates to determine the proportion of children that are likely to make the required developmental transition. Whatever proportion remained would provide an estimate of the prevalence of 'true' dyslexia. Unfortunately the data presented here are insufficient for that purpose.

A widely reported feature of reading disabilities is that it affects a disproportionate number of boys compared to girls. For example in a study of 1206 nine and ten year olds Lewis and colleagues identified more boys than girls as having specific reading disabilities (Lewis et al., 1994). However this apparent gender bias has been challenged by Share and Silva (Share and Silva, 2003). In their study the preponderance of boys identified as having specific reading disabilities was shown to be an artefact of the IQ-discrepancy methodology resulting from differences in the distribution of reading scores between boys and girls. As with the present study the reading scores for girls were found to have a higher mean and smaller variance than those of the boys. When Share and Silva

applied the IQ-discrepancy methodology separately to each gender they actually identified slightly more girls than boys as having specific reading disabilities, 7.7% compared to 6.8%. In this study the results suggest that the rate of reading disabilities is about twice as high in boys as it is in girls in any one year group, but that this results from a developmental lag between boys in girls. Given that the prevalence rate was estimated at 15% for girls in P5 and 16% for boys in P6 it would suggest that this lag is about a year. There is no evidence for any difference between boys and girls in absolute rates of dyslexia.

5. Mathematics

In cohorts P4, P5 and P6 the binormal distribution provided an excellent fit for the observed data leaving little unexplained variation or evidence for additional complexity in the data. As discussed previously the data for the P7 cohort appear to indicate a boost in performance for a select group of children that may have resulted from preparation for grammar school entrance tests. For that reason the P7 data will not be considered further in this discussion.

Tables 9.3.1 and 9.3.2 summarise the parameter estimates for mathematics that were originally presented in table 7.3.

Year Group	**Variable**				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\bar{D}}$	$s_{\bar{D}}$
P4	0.29	-0.72	1.13	0.25	0.97
P5	0.37	-0.77	1.28	0.44	0.81
P6	0.29	-0.84	1.52	0.20	0.88
P7	0.79	-0.19	1.56	1.23	0.76

Table 9.3.1: Summary of binormal parameter estimates for boys' mathematics scores

Year Group	**Variable**				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\bar{D}}$	$s_{\bar{D}}$
P4	0.05	-1.89	0.71	0.07	0.91
P5	0.31	-0.76	1.19	0.36	0.76
P6	0.28	-0.77	1.42	0.16	0.79
P7	0.65	-0.32	1.45	1.00	0.86

Table 9.3.2: Summary of binormal parameter estimates for girls' mathematics scores

Given that the binormal distribution model fits the maths data so well it is initially disappointing that there is no apparent trend in the value of parameter estimates across cohorts. This would suggest that, unlike reading, there is there is no particular evidence for qualitatively separate subtypes of mathematician. This is perhaps not surprising when the nature of the mathematics assessment is considered. Successful engagement with the assessment requires a broader range of skills than does the reading assessment. At the very least it requires a degree of both literacy and numeracy. It is probably also more directly sensitive to the

nuances of the curriculum than is the reading assessment, for example through the use of subject specific vocabulary.

The multi-factorial nature of the mathematics assessment would predict a normal distribution scores. Why then is there such strong evidence that the data is in fact negatively skewed? It would seem to indicate a ceiling effect that has restricted the potential of more able mathematicians to achieve higher scores. An inspection of the probability histograms for mathematics (figures 3.3.1.1 to 3.3.3.2) shows that the weakest mathematicians achieve scores about four years below the average for their age, and yet the most able mathematicians are only two and a half years ahead of the average. The InCAS assessment is capable of providing reliable age-equivalent scores up to at least 16 years, around five years ahead of the oldest participant in this study, so why the ceiling at two-and-half years. If there is no ceiling on the assessment then the next most likely explanation is that there is a ceiling in the curriculum delivery. In order to achieve a score that was four years ahead of the cohort average age a child in the upper primary school would need to have been exposed to the secondary level curriculum. There are any number of reasons why this might not happen. It could be down the confidence and competence of primary teachers to deliver mathematics instruction at such a high level. It might also result from policy decisions concerning curriculum implementation at the school, local authority or national level. If this is correct then it goes some way to explaining why the speculated reason for the boost in P7 scores has had such a marked effect.

6. Arithmetic

At first sight the pattern of results seen in arithmetic was similar to that described for mathematics in the previous section. In general the binormal model provided an excellent fit for the data with some evidence for the emergence of a subpopulation that had received a boost in P7. However there appears to be more evidence of a pattern in the binormal parameter estimates. Tables 9.4.1 and 9.4.2 summarise the parameter estimates for arithmetic that were originally presented in table 7.4.

Year Group	Variable				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\overline{D}}$	$s_{\overline{D}}$
P4	0.19	-1.48	1.51	0.33	0.95
P5	0.22	-1.22	1.72	0.41	0.96
P6	0.27	-0.98	1.61	0.41	0.98
P7	0.24	-0.95	2.06	0.31	1.29

Table 9.4.1: Summary of binormal parameter estimates for boys' arithmetic scores

Year Group	Variable				
	ρ	$\bar{x}_D$	s_D	$\bar{x}_{\overline{D}}$	$s_{\overline{D}}$
P4	0.18	-1.68	1.47	0.14	0.79
P5	0.25	-0.99	1.42	0.33	0.78
P6	0.24	-0.94	1.70	0.26	0.81
P7	0.28	-0.77	1.63	0.31	1.29

Table 9.4.2: Summary of binormal parameter estimates for girls' arithmetic scores

The emergence of a pattern in the binormal parameter estimates for arithmetic that was not apparent in the mathematics data may be a direct consequence of the relative simplicity of the assessment task. Successful interaction with the InCAS arithmetic assessment depends on a narrower range of cognitive skills than are required for mathematics. This will result is a less complex data structure, and therefore a greater chance that the binormal model will reveal meaningful consistencies in the data.

The trend in the estimate of the mean score of the low attaining subpopulation to increase with the age of the cohort suggests that the arithmetical skills of the low attaining children is catching up with those of the higher attaining children. However The InCAS arithmetic assessment is restricted to a relatively simple format of items. In consequence there is a known ceiling on the assessment overall of about 14 years, whilst the ceiling on the addition subtest is as low as 11 years. The observed pattern probably reflects the lack of capacity of the assessment to extend the more able children. This ceiling effect would also explain the negative skew observed in the score distributions, which is in contrast to the ceiling in curriculum delivery that was proposed in the case of mathematics.

Even if the assessment has a ceiling for able arithmeticians this should not be sufficient to affect the scores of a subpopulation with specific arithmetical disabilities. However it is observed that the standard deviation estimates for the low attaining group are very high, often exceeding that observed for the whole population and reported in table 2.4.2. Whilst it is acknowledged that it is not necessarily so, it might be expected that a qualitatively distinct group of weak arithmeticians would show less variation in the distribution of their scores. If this is coupled with an expectation that the prevalence of such a subpopulation be considerably lower than that estimated here (between 18% and 28%), then an alternative explanation for the pattern of results seems more plausible. It seems likely that in these circumstances the limited flexibility of the binormal model has been utilised to explain the skew caused by the ceiling in the assessment. However it might still be possible to reveal a group with specific arithmetical disabilities if a higher modality model, such as a trinormal model were employed.

Our current understanding of the nature of specific arithmetical disabilities would suggest there may be several subtypes weak arithmeticians, and that a high modality model may be necessary to reveal them. The medically equivalent term for arithmetical learning disabilities is dyscalculia, a condition that was originally proposed by Kosc (Kosc, 1970, Kosc, 1974). There are two competing hypotheses for the underlying cognitive deficit responsible for dyscalculia (Feigenson et al., 2004). Butterworth has proposed the *defective number module hypothesis* (Butterworth, 2005b). According to Butterworth the fault lies with a deficit in the innate ability to understand and manipulate small whole number quantities. The competing hypothesis proposed by Deheane and colleagues states that the deficit lies with the cognitive systems involved in magnitude representation and which allow us to understand approximate quantities (Dehaene et al., 2004, Dehaene et al., 2003, Wilson and Dehaene, 2007). Dehaene calls this ability number sense. Since these models are not mutually exclusive there is the possibility of two distinct categories of arithmetical disability based on this theory alone, but the complications do not end there.

One of the difficulties with studying arithmetical learning disabilities is its apparent association with so many other conditions such as working memory deficits, ADHD and dyslexia (von Aster and Shalev, 2007). This has led Rubinsten and Henik to propose three alternative frameworks for the classification of arithmetical learning disabilities according the hypothesised cognitive deficits underlying the condition. These range from the shared deficits that underlie other conditions such as dyslexia to those that are very specific. According to their model the term dyscalculia should be restricted to cases where the causal cognitive deficit lies with the processing numerical quantities alone (Rubinsten and Henik, 2009). Whilst arithmetical and reading disabilities are often reported as being comorbid there is increasing evidence for a dissociation between dyscalculia and dyslexia, therefore suggesting different underlying cognitive deficits for the two conditions (Landerl et al., 2009, Rubinsten and Henik, 2006, Swanson and Jerman, 2006).

Although prevalence estimates for arithmetical learning disabilities have been put in the region of 5% to 6% it is likely that cases of 'pure' dyscalculia are very rare indeed (Snowling, 2005, Gifford and Rockliffe, 2008). In a study of 1206 nine and ten year olds Lewis and colleagues reported that the prevalence of specific arithmetic difficulties was three times lower than that of specific reading difficulties (Lewis et al., 1994). The figures for arithmetic and reading were put at 1.3% and 3.9% respectively. Given the low expected prevalence and multiplicity of possible subtypes of specific arithmetical disabilities finding evidence for their existence, even using an enhanced multimodal version of the method described here, may present considerable challenges. However, if viewed from a developmental rather than neurocognitive perspective the potential of the methodology may be more promising.

In the analysis of reading data presented in section four of this chapter evidence was presented for a developmental step that may represent the transition from beginning to fluent reader. Theoretically it may be possible to find a parallel transition in the case of arithmetic. The mechanism by which children acquire arithmetical skills is well understood, at least in the case of addition (Butterworth, 2005a, Geary, 2003, Geary and Hoard, 2005). Initially children

perform arithmetic using cumbersome counting strategies. As their skills improve they gradually adopt more efficient calculation techniques. With enough practice they may eventually commit a number fact to long-term memory and will therefore be able to instantly recall the answer to that particular arithmetical problem. This change from a calculation to a recall strategy represents a fundamental shift that may be detectable in the data. It has been shown that the tendency to progress from calculation to recall strategies does not occur evenly across arithmetical operations. For example recall of the solution to addition problems is far more likely than it is for subtraction (Barrouillet et al., 2008). It has also been shown that whilst the rate of recall of multiplication facts tends to increase with the age of the child, division facts are rarely committed to memory (Robinson et al., 2006, Steel and Funnell, 2001). Clearly if evidence for a developmental shift from calculation to recall strategies is to be discovered in the InCAS arithmetic scores it will be necessary to analyse the data from the four subtests separately.

7. Summary

In this chapter the validity of the binormal model fits generated in this study have been evaluated. The rationale for developing a methodology for deriving binormal parameter estimates was to discover evidence for the existence of qualitatively distinct subpopulations of learner in the population. It was argued that if found it would provide evidence to support the medical model of learning disabilities that was largely free of the criticisms that have been widely levelled at traditional IQ-discrepancy based methodologies. In practice no evidence was found for neurocognitive deficits that would indicate the existence of a specific dysfunction such as dyslexia or dyscalculia. However evidence was found for a developmental transition at around the age of seven years in the case of reading. Evidence was also found for a developmental delay of about a year between the genders that would account for the higher proportion of boys with reading disabilities that are reported in many studies. Although no evidence was found that would support the existence of a specific neurocognitive deficit it was argued that ‘true’ dyslexics might represent that portion of the population that fail to make the transition, and that evidence for this might still be found by

looking at the trend in prevalence estimates across a suitable longitudinal or cross-sectional dataset.

Evidence was presented that the methodology is sensitive to ceiling effects in both assessment and curriculum delivery. It was also argued that the binormal model was of limited use when applied to datasets with a complex structure, specifically those with a modality higher than two. Two approaches were suggested to counter this difficulty. The first approach would be to develop higher modality models such as the trinormal distribution. The second approach would be to apply the existing model in the case of assessments designed to measure more specific abilities, for example addition as opposed to arithmetic.

Chapter 8: Final Discussion

1. Critique of the Study

The principal weakness of the study is that whilst it has provided a unique perspective on the nature of learning disabilities and given fresh insights, the strength of any conclusions that might be drawn from the analyses are tempered by limitations of the methodology employed for deriving binormal parameter estimates. These limitations are discussed below.

It was argued in chapter 3 that successful application of the methodology is dependent on both the quality and quantity of the data available. It was then shown in chapter 4 that it is necessary to be mindful of factors that might affect the distribution of assessment scores. Specifically the effect of gender and years of schooling were taken into consideration by analysing these data separately. In addition the effect of intra-cohort age differences was taken into account by performing the analysis on the age-ability difference scores, rather than on the assessment data directly.

Having anticipated and accounted for gender, years of schooling and intra-cohort age differences, the methodology revealed little evidence for qualitatively distinct subtypes of learner as it was intended to do. The best evidence for this was found in the data for reading, but the pattern of results across cohorts was inconsistent. It was argued that the methodology was sensitive to additional complexity in the data that might be introduced by ceiling (and presumably floor) effects, instructional effects and multiplicity in the constructs that a particular assessment was designed to measure. With respect to ceiling effects evidence was found for an assessment ceiling in the case of arithmetic, and an instructional ceiling in the case of mathematics. It was argued that the ceiling resulted in a skew in the distribution of scores for the higher achieving subpopulation, and that the limited flexibility in the binormal model was used to describe that skewness, rather than any bimodality that may have been present in the data. Evidence for further instructional effects came from the apparent boost in arithmetic and particularly mathematics scores in the final year of primary school

(P7). It was argued that this boost may have resulted from the preparation of select individuals for secondary school entrance tests because this may have increased curriculum knowledge or contributed to general test-wiseness. The effect of construct multiplicity was illustrated by the picture vocabulary assessment where it was argued that the results were so influenced by the effect of reading on vocabulary development that it failed to reveal qualitatively different subtypes of English language speaker.

A final limitation in the methodology that is inherent in the nonlinear regression technique is the danger of arriving at misleading model fits, particularly if the model uses a number of variable parameters. One way to approach this difficulty is to evaluate any model fit for validity within a theoretical framework. Another approach is to look for trends and patterns in model fits across different datasets. Validation of the methodology presented here will ultimately depend upon the establishment of explainable patterns across diverse datasets. In this study within-population patterns have been sought within a cross-sectional dataset. The methodology would benefit from application to a within-population longitudinal dataset, and also a cross-population analysis.

In spite of these limitations the study has revealed some interesting results, particularly in the case of reading. The issues highlighted here are not insurmountable and the methodology stands as a proof of concept. With further refinement and extension of the methodology there is potential for it to shed light on the nature of learning disabilities that is free of the dogma of IQ-discrepancy.

2. Refinement and Extension of the Methodology

The method for deriving binormal parameter estimates described in chapter 3 uses two constants and three variables to define the five binormal parameters. The constants employed are the mean and variance of the distribution. A desirable refinement to the methodology would be to substitute one or more of the remaining variables with constants, the value of which could be calculated directly from the data. To do so would reduce the chance of obtaining multiple model fits.

The variance represents a specific example from a family of distribution shape statistics with the general form:

$$\sigma^a = \frac{\sum(X-\mu)^a}{n}$$

In the case of the variance the value of $a = 2$ is substituted into the expression. When a value of $a = 3$ is used the expression provides a measure of the asymmetry in the data that is akin to the skew. In the case of a symmetrical distribution such as the normal distribution the value of $\sigma^3 = 0$. In principle it is possible to expand this expression along the same lines as that employed in chapter 2, and so derive an expression for the asymmetry of the binormal distribution in terms of the binormal parameters. In practice the algebra required for this is quite complicated. However a 3-constant / 2-variable model has been implemented in DataFit in the case of the standardised binormal distribution. This was possible because the substitution of population mean and variance figures of 0 and 1 respectively made the calculations considerably simpler.

A serious limitation of the methodology is its inability to cope with more complex data structures such as a trimodal distribution, or a bimodal distribution in which one of the subpopulation distributions is skewed. A solution to this is to fit a trinormal distribution to the data instead. The probability density function of the trinormal distribution in variate X is given by:

$$P(X) = \rho_D P(X_D) + \rho_{\bar{D}} P(X_{\bar{D}}) + (1 - \rho_D - \rho_{\bar{D}}) P(X_{\tilde{D}})$$

Adding a third subpopulation increases the number of parameters required to describe the model by three to a total of eight. The extra parameters are needed to describe the prevalence, mean and standard deviation of the additional subpopulation. A drawback of this is that increasing the number of variable parameters also increases the chance of finding multiple model fits. Clearly application of such a model would require even greater care than is the case with the binormal model.

In principle the modality of the model could be expanded still further, but the number of parameters required to describe the model would increase by three with each additional subpopulation. Ideally models of increasing modality would be applied to the data as long as the value of the adjusted coefficient of multiple determination (Ra^2) continued to increase, with initial parameter estimates informed by the solution arrived at from the application of the preceding model.

3. Diagnostic Utility

The binormal subpopulation plots presented in figure 4 graphically illustrate a problem at the heart of the identification of learning disabilities. The overlap between the low attaining subpopulation and the weaker members of the higher attaining subpopulation means that it is impossible to accurately assign every individual to the correct subpopulation based upon their assessment score alone. Many studies assign individuals to one category of learning disability or another based on a particular cut-score. However it is inevitable that whatever cut-score is chosen some individuals will be wrongly assigned.

In medicine incorrect diagnosis can have serious consequences, and so much effort has been expended in the development of protocols for establishing the diagnostic utility of screening tests. One such protocol is the use of receiver operating characteristic (ROC) curves (Park et al., 2004, Faraggi and Reiser, 2002). An ROC curve is that which results when the true positive fraction (TPF) is plotted against the false positive fraction (FPF) for a range of cut-scores on a particular screening test. The true positive fraction is defined as the proportion of individuals below the cut-score that were correctly identified as having the particular disease or condition. The false positive fraction is defined as the proportion of individuals above the cut score that were incorrectly identified as not having the disease or condition. If the distributions of diseased and healthy individuals each follow a normal distribution, that is if the scores are binormaly distributed in the population, it can be shown that the ROC curve is described by the following expression where Φ represents the standard normal cumulative distribution function:

$$TPF = \Phi\left(\left(\frac{\bar{x}_D - \bar{x}_{\bar{D}}}{s_D}\right) - \left(\frac{s_{\bar{D}}}{s_D}\right)\Phi^{-1}\left(FPF\right)\right)$$

This expression has been adapted from the one reported by Park and colleagues to utilise the notation of Pepe that has been used throughout this thesis (Park et al., 2004, Pepe, 2003)

The area under this curve can take any value between 0.5 and 1. A value of 0.5 for the area under the curve would indicate that the screening test had no power to correctly assign individuals to the correct category of either diseased or healthy. A value of 1 would indicate that the assessment correctly assigns individuals every time. Thus the magnitude of the area under the ROC curve is measure of the diagnostic utility of the screening test.

Clinicians use the data contained in ROC curves to inform the judgements they make as to the most appropriate cut-score to use. In making those judgements it is necessary to weigh up the consequences of treating those individuals with an incorrect positive diagnosis against not treating those with an incorrect negative diagnosis. Clearly the same type of information would have a similar value for informing the decisions made by teachers, psychologists and researchers in the field of learning disabilities.

To illustrate one such application of ROC analysis the diagnostic utility of the InCAS reading assessment will be considered for girls in cohorts P4, P5 and P6. These data were selected because they represent the strongest evidence found in the present study for qualitatively distinct subtypes of learner. The ROC curves on which the following analysis is based are presented in figure 5.

For data that are binormaly distributed the area under the ROC curve (AUC) can be calculated using the expression below which has been adapted from the one reported by Faraggi and Reiser to utilise the notation of Pepe (Faraggi and Reiser, 2002, Pepe, 2003):

$$AUC = \Phi\left(\frac{\bar{x}_{\overline{D}} - \bar{x}_{D}}{\sqrt{\left(s_{\overline{D}}^2 + s_{D}^2\right)}}\right)$$

Inputting the derived binormal parameters reported in tables 7.2 into this equation produced the following values for the area under the curve; 0.78 in P4, 0.88 in P5 and 0.97 in P6. The figures indicate that the diagnostic utility of the InCAS reading assessment to correctly categorise the children into the low attaining or high attaining subtype increases with the age of the cohort. Even though the prevalence of the low attaining readers decreases, the chance of correctly identifying them increases.

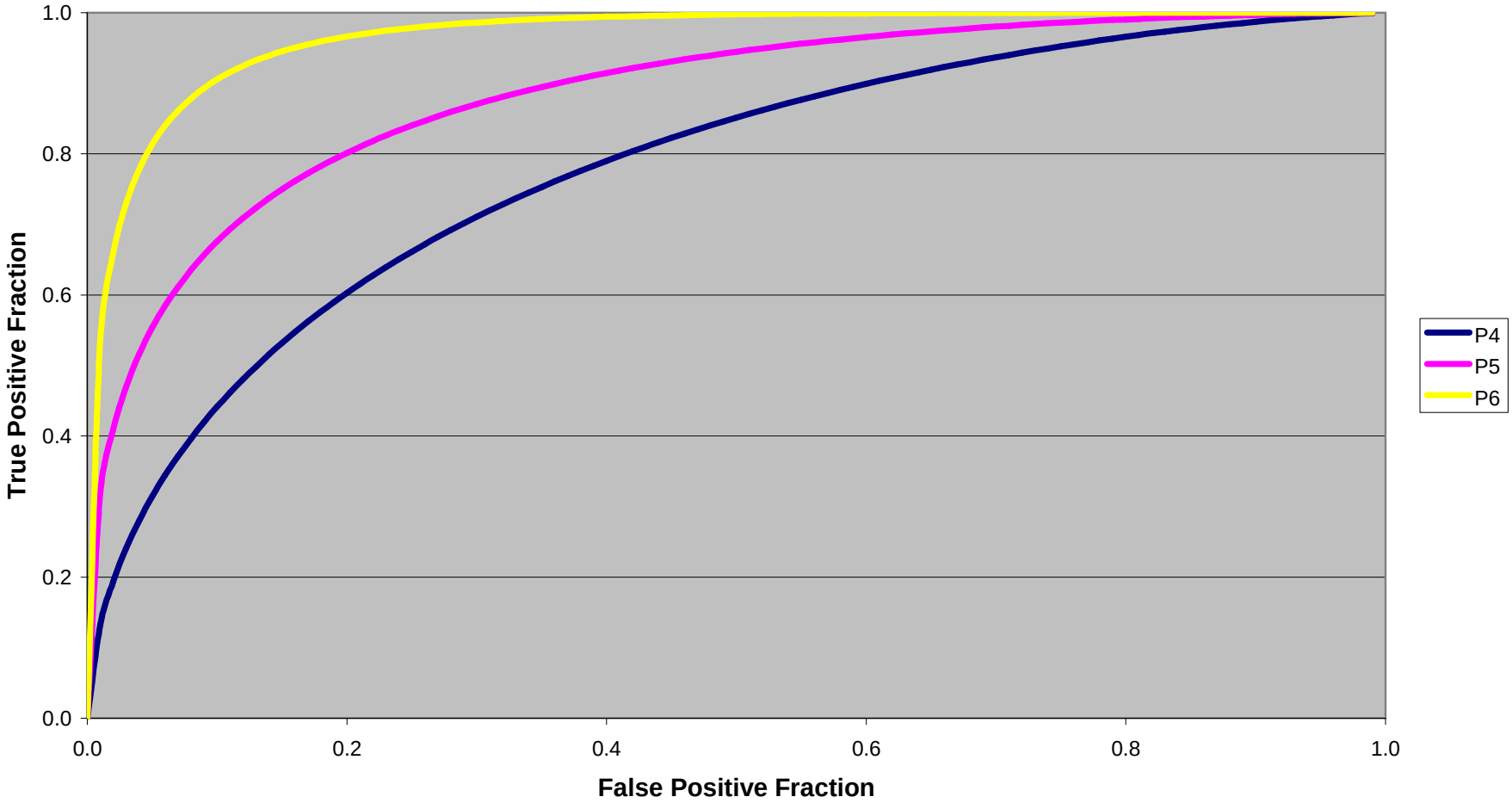
The ROC methodology clearly has potential for rating the diagnostic utility of an assessment and informing choice of cut-score. In the field of learning disabilities research it would allow researchers to estimate and report the proportion of children in their treatment groups that had been incorrectly classified.

A related application of the binormal distribution model lies in the establishment of the chance that an individual belongs to one or other subtype of learner. Rather than categorising a child as having a learning disability according to which side of a cut-score their assessment result fell it would be possible to calculate a probability that they fell into the learning disability group. This approach might be particularly useful when coupled to longitudinal monitoring of children.

Clearly the application of the binormal model of specific learning disabilities has a great potential to provide an additional dimension to the interpretation of data generated by screening tests. However these potential applications are dependent on establishing the validity of the groupings revealed by the binormal model fitting procedure. The ROC approach also highlights the fact that such assessments are unlikely to ever be completely accurate in their designation of learning disabilities. However reliable and valid an adaptive assessment such as

InCAS may be, the computer remains an unintelligent observer. Ultimately the diagnosis of learning disabilities requires intelligent observation.

Figure 5: ROC Curve for Girls' Reading Results



4. Conclusions

In conclusion to the thesis the research questions posed in chapter 1 will now be reconsidered. These questions are reproduced below:

1. Does the binormal distribution provide a suitable model for the investigation of bimodality in an epidemiological study of academic attainment in primary school children?
2. Is there any evidence for qualitatively distinct subtypes of learner in the population under study?
3. Is it possible to obtain valid and reliable parameter estimates for the distribution of assessment scores for different subtypes of learner within the population as a whole?
4. To what extent does the identification of distinct subtypes of learner support the medical model of specific learning disabilities? Is there any evidence for the existence of dysfunctions such as dyslexia and dyscalculia?
5. What are the implications of application of the binormal model to the identification of children with specific learning disabilities?

In the present study the evaluation of 32 datasets defined by cohort, gender and InCAS assessment module was undertaken. It was found that in all but one case the binormal distribution model provided a better description of the distribution of age-ability difference scores than did the simpler normal distribution model. However real evidence of bimodality was only found in four cases; that is the reading data for girls in P4 and P5, and the reading data for both boys and girls in P6. In other circumstances additional flexibility afforded by the binormal model was utilised to explain other structure within the score distributions such as skew and higher order modality. Nevertheless it was argued that the application of the binormal model stands as a proof of concept. It was also proposed that the utility of the general methodological approach introduced here might be extended by developing higher modality models or else by applying the binormal distribution model to assessments that are designed to measure more specific cognitive functions.

The evidence for qualitatively distinct subtypes of learner was restricted to some reading assessments, as stated in the previous paragraph. The pattern in the data suggested that the qualitative difference between the identified subpopulations may represent a developmental step that occurs at a mean age of about seven years. In this context reading disabilities would be manifested in a developmental lag. Evidence for a developmental lag of about one year between boys and girls was observed. It was speculated that the developmental step may represent the transition from beginning to fluent reader, although it was acknowledged that further research would be required to validate that idea.

The methodology employed certainly makes it possible to obtain reliable binormal parameter estimates. The nonlinear regression technique generates standard error statistics on the magnitude estimates of variable parameters in the model. These were used to evaluate whether or not each variable parameter makes a significant contribution to the model fit. However the production of reliable statistics does not imply that they have any validity. In order to establish validity cross-cohort and cross-gender patterns were sought in the binormal parameter estimates. These patterns were then interpreted within a theoretical framework. This approach yielded positive results in the case of some reading assessments as discussed above.

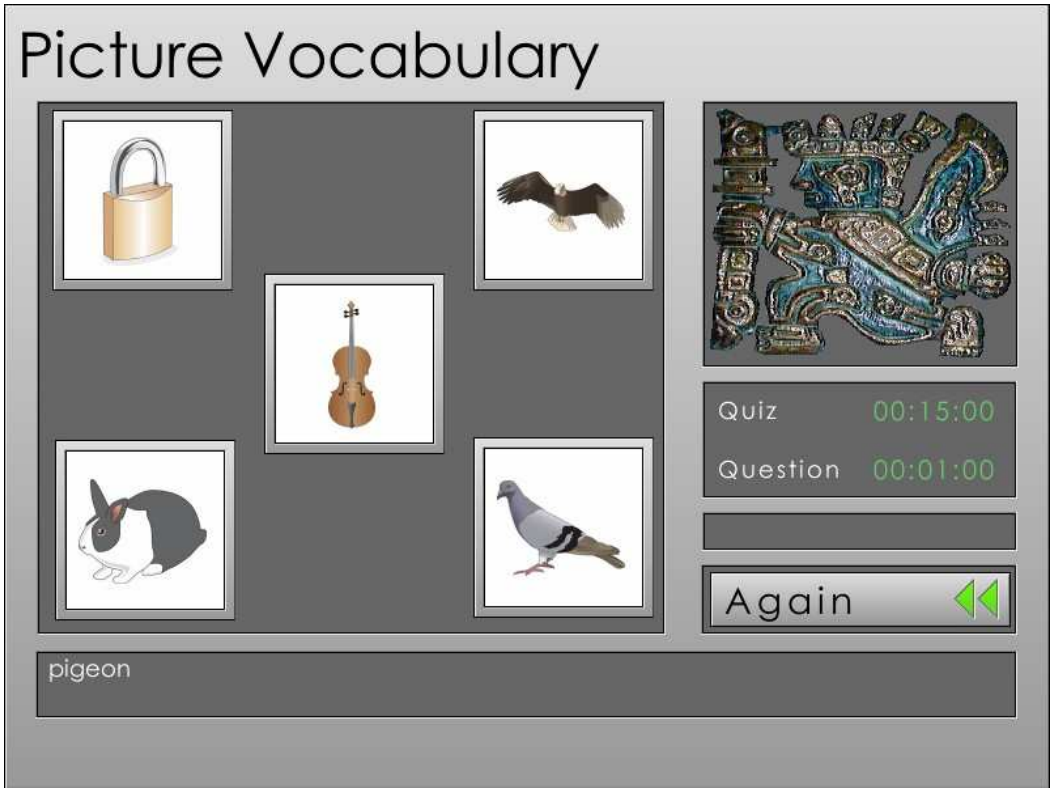
The evidence for the existence of distinct subtypes of reader was interpreted within a behavioural rather than medical framework. The medical model predicts a fixed prevalence of dysfunction across cohorts, but what was observed was a decrease in the prevalence of low ability readers from one cohort to the next. The decrease was observed to be in the region of 50% each year. The prevalence of low attaining readers was also higher than might be expected from reported prevalence estimates of dyslexia. If all children were to ultimately make the transition from the low to high attaining group of readers it would support the hypothesis that reading abilities represent a single continuum in the population. It was proposed that true dyslexics might be composed of that group of children that ultimately fail to make the transition, but the data were insufficient to determine if such a group were likely to exist. No evidence was found to support

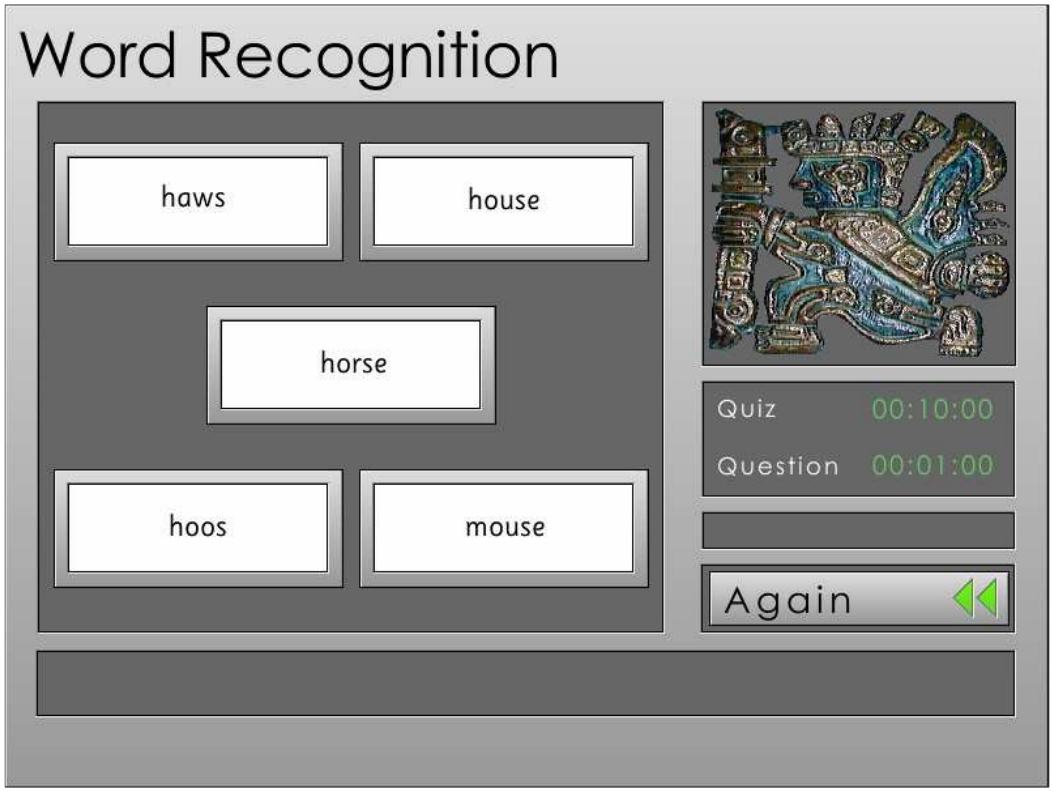
the existence of dyscalculia, but this may simply result from limitations of the methodology and the low expected prevalence of the condition.

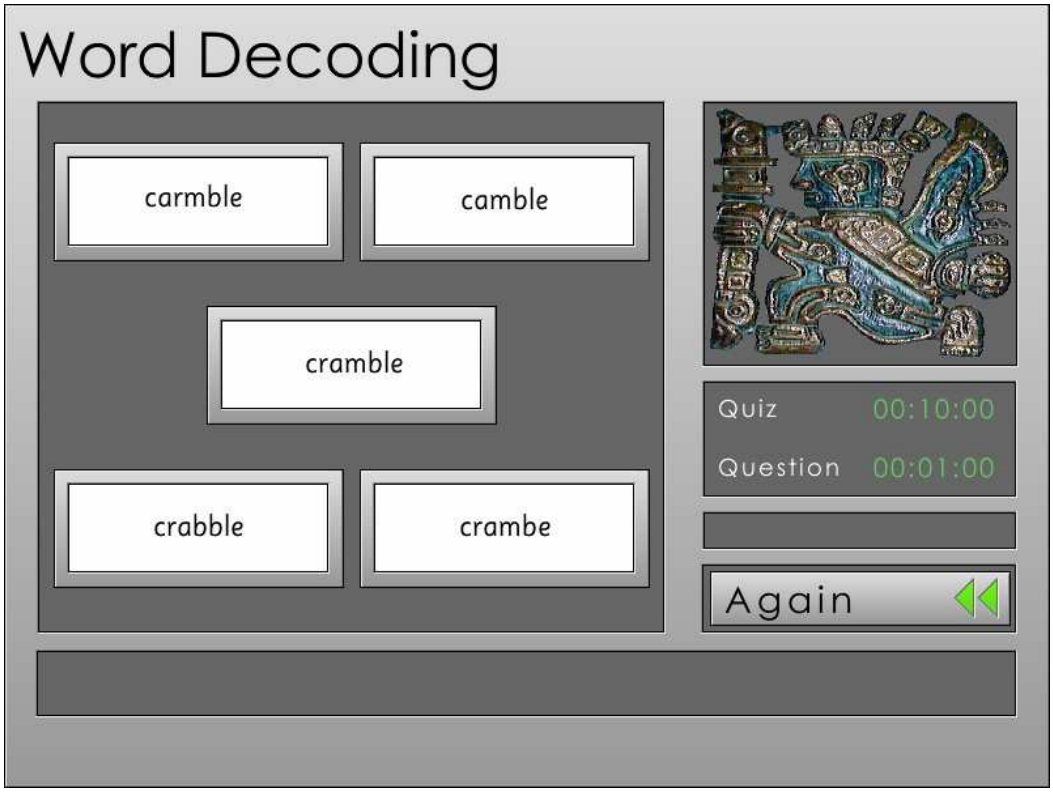
Finally, the application of the binormal model to define children in terms of the probability that they have a learning disability was discussed. At the group level this would allow estimates to be made of the rate of misdiagnosis on the basis of cut-scores. It could also be used to inform the best choice of cut-score and to establish the diagnostic utility of a screening test. At the individual level it would provide additional perspective to teachers and clinicians in the interpretation of assessment data.

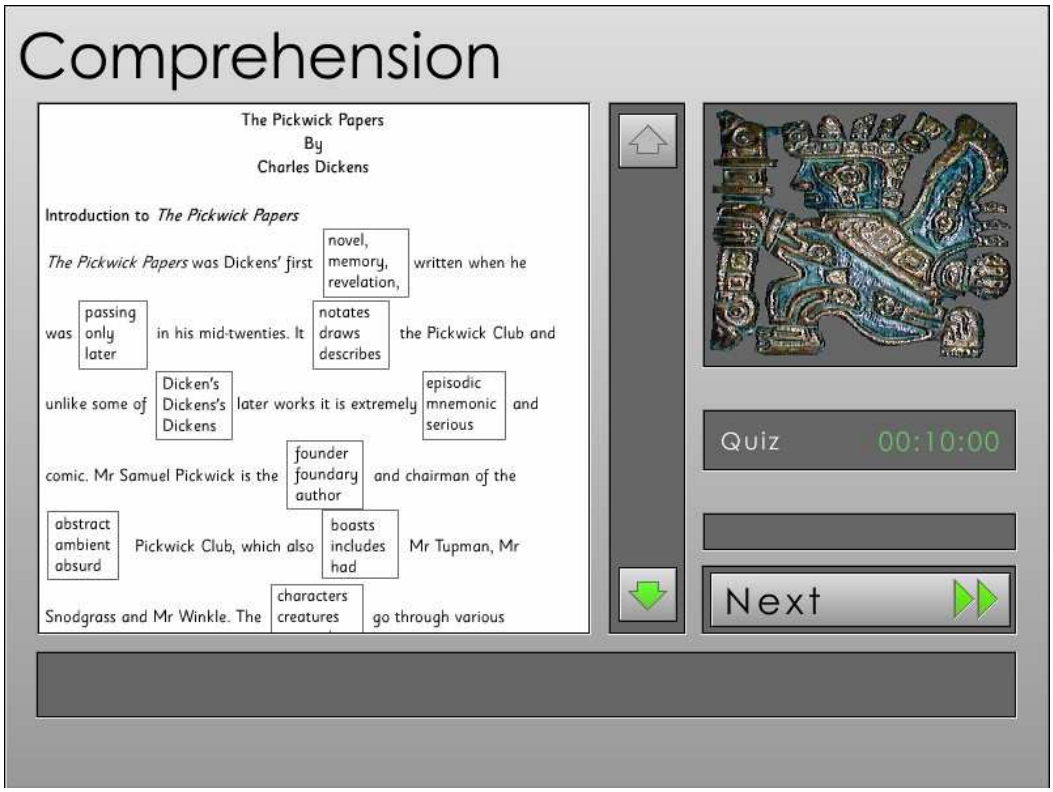
Despite the acknowledged limitations of the methodology the binormal modelling approach offers real potential to give new insights into the nature of learning disabilities. It presents a perspective that is free of the widely reported difficulties associated with the IQ-discrepancy technique. However its real strength lies with the complete objectivity of the parameter estimation process. To the authors knowledge all previous research in the field of learning disabilities has included some subjective element, even if it has only been in the choice of cut-score. Of course interpretation of the output from any analytical procedure will always have some subjectivity. The difficulties associated with the methodology are not insurmountable and the rewards for doing so may be considerable.

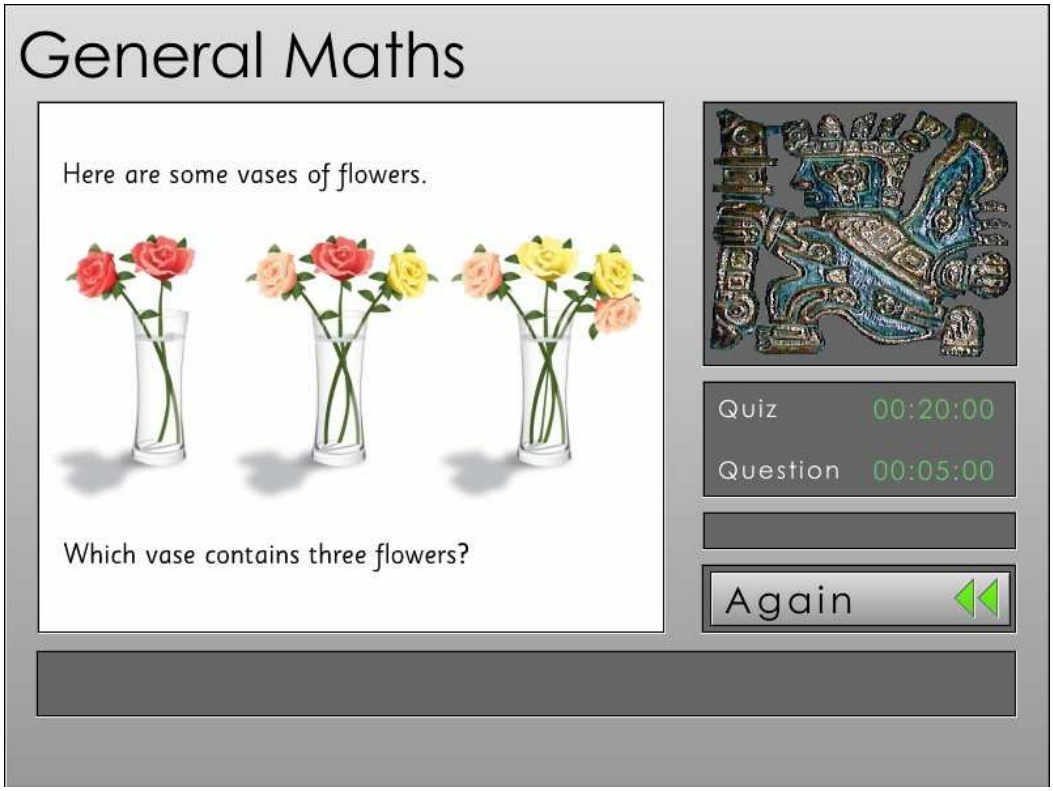
Appendix: Example Screenshots of the InCAS Assessment Modules

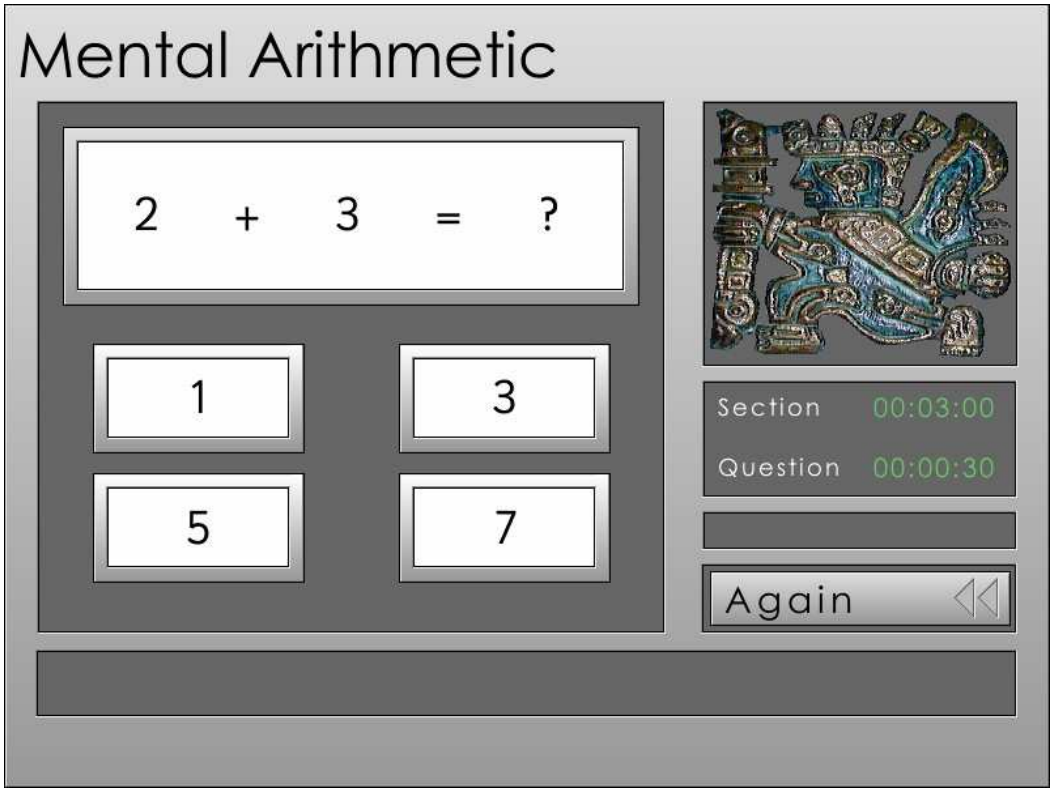





Word Decoding
carmble
camble
cramble
crabble
crambe
Quiz 00:10:00
Question 00:01:00
Again


Comprehension
The Pickwick Papers
By
Charles Dickens
Introduction to The Pickwick Papers
The Pickwick Papers was Dickens' first novel, memory, revelation, written when he
was passing only later in his mid-twenties. It notates draws describes the Pickwick Club and
unlike some of Dicken's Dickens's Dickens later works it is extremely episodic mnemonic serious and
comic. Mr Samuel Pickwick is the founder foundary author and chairman of the
abstract ambient absurd Pickwick Club, which also boasts includes had Mr Tupman, Mr
Snodgrass and Mr Winkle. The characters creatures go through various
Quiz 00:10:00
Next


General Maths
Here are some vases of flowers.
Which vase contains three flowers?
Quiz 00:20:00
Question 00:05:00
Again


Mental Arithmetic
2 + 3 = ?
1
3
5
7
Section 00:03:00
Question 00:00:30
Again

References

AARON, P. G. (1997) The impending demise of the discrepancy formula. *Review of Educational Research,* 67**,** 461-502.

ALBONE, S., TYMMS, P. & ADAMS, J. W. (2006a) The diagnostic utility of PIPS standardised reading assessments. *European Association for Research on Learning and Instruction (EARLI) Conference.* Northumbria University, UK.

ALBONE, S., TYMMS, P. & ADAMS, J. W. (2006b) Specific learning disabilities: An objective method for the determination of prevalence estimates. *British Educational Research Association Annual Conference.* University of Warwick, UK.

ANDERSON, P. L. & MEIER-HEDDE, R. (2001) Early case reports of dyslexia in the United States and Europe. *Journal of Learning Disabilities,* 34**,** 9-21.

BARROUILLET, P., MIGNON, M. & THEVENOT, C. (2008) Strategies in subtraction problem solving in children. *Journal of Experimental Child Psychology,* 99**,** 233-251.

BERLIN, R. (1884) Über Dyslexie. *Archiv für Psychiatrie,* 15**,** 276-278.

BERLIN, R. (1887) *Eine besondere Art der Wortblindheit,* Wiesbaden, Germany, J. F. Bergman.

BOND, T. G. & FOX, C. M. (2001) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.,* Mahwah, N.J., Lawrence Erlbaum Associates.

BUTTERWORTH, B. (2005a) The development of arithmetical abilities. *Journal of Child Psychology and Psychiatry,* 46**,** 3-18.

BUTTERWORTH, B. (2005b) Developmental Dyscalculia. IN CAMPBELL, J. I. D. (Ed.) *The Handbook of Mathematical Cognition.* New York, Psychology Press.

COHEN, L. (1979) Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology,* 32**,** 113-20.

CRITCHLEY, M. (1970) *The dyslexic child,* London, Heinemann Medical.

DEHAENE, S., MOLKO, N., COHEN, L. & WILSON, A. J. (2004) Arithmetic and the brain. *Current Opinion in Neurobiology,* 14**,** 218-224.

DEHAENE, S., PIAZZA, M., PINEL, P. & COHEN, L. (2003) Three parietal circuits for number processing. *Cognitive Neuropsychology,* 20**,** 487-506.

EVERITT, B. S. (1981) Bimodality and the nature of depression. *British Journal of Psychiatry,* 138**,** 336-339.

FARAGGI, D. & REISER, B. (2002) Estimation of the area under the ROC curve. *Statistics in Medicine,* 21**,** 3093-3106.

FEIGENSON, L., DEHAENE, S. & SPELKE, E. S. (2004) Core systems of number. *Trends in Cognitive Sciences,* 8**,** 307-314.

FLEISS, J. L. (1972) Classification of the depressive disorders by numerical typology. *Journal of Psychiatric Research,* 9**,** 141-153.

FOERSTER, R. (1905) Beiträge zur pathologie des lesens und schreibens (congenitale wortblindheit bei einem schwachsinnigen). *Neurologisches Zentralblatt,* 24**,** 235.

GEARY, D. C. (2003) Learning Disabilities in Arithmetic: Problem-Solving Differences and Cognitive Deficits. IN SWANSON, H. L., HARRIS, K. R. & GRAHAM, S. (Eds.) *Handbook of Learning Disabilities.* New York, Guildford Press.

GEARY, D. C. & HOARD, M. K. (2005) Learning Disabilities in Arithmetic and Mathematics: Theoretical and Empirical Perspectives. IN CAMPBELL, J. I. D. (Ed.) *The Handbook of Mathematical Cognition.* New York, Psychology Press.

GIFFORD, S. & ROCKLIFFE, F. (2008) In search of dyscalculia. IN JOUBERT, M. (Ed.) *Proceedings of the British Society for Research into Learning Mathematics.* University of Manchester.

GOUGH, P. B. & TUNMER, W. E. (1986) Decoding, reading, and reading disability. *Remedial and Special Education,* 7**,** 6-10.

GRINSTEAD, C. M. & SNELL, J. L. (1997) *Introduction to probability,* American Mathematical Society.

GRUENBERG, E. (1966) Epidemiology of mental illness. *International Journal of Psychiatry,* 2**,** 78-134.

HINSHELWOOD, J. (1896) A case of dyslexia: A peculiar form of word blindness. *The Lancet,* 148**,** 1451-1454.

HINSHELWOOD, J. (1900) Congenital word-blindness. *The Lancet,* 155**,** 1506-1508.

HOPKINS, K. D. (1998) *Educational and Psychological Measurement and Evaluation,* Boston, Allyn and Bacon.

IEA (1995a) TIMMS Mathematics Items: Released Set for Population 1 (Third and Fourth Grades). International Association for the Evaluation of Educational Achievement.

IEA (1995b) TIMMS Mathematics Items: Released Set for Population 2 (Seventh and Eighth Grades). International Association for the Evaluation of Educational Achievement.

JIMÉNEZ GONZÁLEZ, J. E. & GARCIA ESPÍNEL, A. I. (1999) Is IQ-achievement discrepancy relevant in the definition of arithmetic learning disabilities? *Learning Disability Quarterly,* 22**,** 291-301.

JORM, A. F., SHARE, D. L., MACLEAN, R. & MATTHEWS, R. (1986) Cognitive factors at school entry predictive of specific reading retardation and general reading backwardness: A research note. *Journal of Child Psychology,* 27**,** 45-54.

KAVALE, K. A. (2001) Discrepancy models in the identification of learning disability. *Office of Special Education Programs (OSEP) Learning Disabilities Summit: Building a Foundation for the Future.* Washington, D.C.

KAVALE, K. A. (2005) Identifying Specific Learning Disability: Is Responsiveness to Intervention the Answer? *Journal of Learning Disabilities,* 38**,** 553-562.

KAVALE, K. A. & FORNESS, S. R. (2000) What definitions of learning disability say and don't say: A critical analysis. *Journal of Learning Disabilities,* 33**,** 239-256.

KENDEOU, P., SAVAGE, R. & VAN DEN BROECK, P. (2009) Revisiting the simple view of reading. *British Journal of Educational Psychology,* 79**,** 353-370.

KERR, J. (1897) School hygiene, in its mental, moral, and physical aspects. *Journal of the Royal Statistical Society,* 60**,** 613-680.

KIRK, S. A. (1962) *Educating Exceptional Children,* Boston, Houghton Mifflin.

KOSC, L. (1970) Contribution to the nomenclature and classification of the disorders in mathematical abilities. *Studia Psychologica,* 12**,** 12-28.

KOSC, L. (1974) Developmental dyscalculia. *Journal of Learning Disabilities,* 7**,** 164-177.

LANDERL, K., FUSSENEGGER, B., MOLL, K. & WILLBURGER, E. (2009) Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experimental Child Psychology,* 103**,** 309-324.

LEWIS, C., HITCH, G. J. & WALKER, P. (1994) The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10-year-old boys and girls. *Journal of Child Psychology and Psychiatry,* 35**,** 283-292.

LINACRE, J. M. (2007) Winsteps version 3.64.0. Chicago, IL.

MERRELL, C. & TYMMS, P. (2007) Identifying reading problems with computer adaptive assessments. *Journal of Computer Assisted Learning,* 23**,** 27-35.

MORGAN, W. P. (1896) A case of congenital word blindness. *British Medical Journal,* 2**,** 1378.

MORTON, J. & FRITH, U. (1995) Causal modeling: A structural approach to developmental psychopathology. IN CICCHETTI, D. & COHEN, D. J. (Eds.) *Manual of developmental psychopathology.* New York, Wiley.

NETTLESHIP, E. (1901) Cases of congenital word-blindness (inability to learn to read). *Ophthalmic Review,* 20**,** 61-67.

OAKDALE ENGINEERING (2008) DataFit version 9.0.59. 9.0.59 ed. Oakdale, PA. USA.

PARK, S. H., GOO, J. M. & JO, C.-H. (2004) Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean Journal of Radiology,* 5**,** 11-18.

PENNINGTON, B. F., GILGER, J. W., OLSEN, R. K. & DEFRIES, J. C. (1992) The external validity of age- versus IQ-discrepancy definitions of reading disability: lessons from a twin study. *Journal of Learning Disabilities,* 25**,** 562-573.

PEPE, M. S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction.*, Oxford, UK, Oxford University Press.

PREECE, P. F. W. (2002) Equal-interval measurement: the foundation of quantitative educational research. *Research Papers in Education,* 17**,** 363-372.

REES, D. G. (1987) *Foundations of Statistics,* London, Chapman and Hall.

ROBINSON, K. M., ARBUTHNOTT, K. D., ROSE, D., MCCARRON, M. C., GLOBA, C. A. & PHONEXAY, S. D. (2006) Stability and change in children's division strategies. *Journal of Experimental Child Psychology,* 93**,** 224-238.

RODGERS, B. (1983) The identification and prevalence of specific reading retardation. *British Journal of Educational Psychology,* 53**,** 369-373.

RUBINSTEN, O. & HENIK, A. (2006) Double dissociation of functions in developmental dyslexia and dyscalculia. *Journal of Educational Psychology,* 98**,** 854-867.

RUBINSTEN, O. & HENIK, A. (2009) Developmental Dyscalculia: heterogeneity might not mean different mechanisms. *Trends in Cognitive Sciences,* 13**,** 92-99.

RUTTER, M. (1978) Prevalence and types of dyslexia. IN BENTON, A. L. & PEARL, D. (Eds.) *Dyslexia: An appraisal of current knowledge.* New York, Oxford University Press.

RUTTER, M. & YULE, W. (1975) The concept of specific reading retardation. *Journal of Child Psychology and Psychiatry,* 16**,** 181-197.

SCHONELL, F. J. (1935) Diagnostic tests for specific disabilities in school subjects. *Year Book of Education.* London, Evans.

SHARE, D. L., MCGEE, R., MCKENZIE, D., WILLIAMS, S. & SILVA, P. A. (1987) Further evidence related to the distinction between specific reading retardation and general reading backwardness. *British Journal of Developmental Psychology,* 5**,** 35-44.

SHARE, D. L. & SILVA, P. A. (2003) Gender bias in IQ-discrepancy and post-discrepancy definitions of reading disability. *Journal of Learning Disabilities,* 36**,** 4-14.

SHAYWITZ, B. A., FLETCHER, J. M., HOLAHAN, J. M. & SHAYWITZ, S. E. (1992a) Discrepancy compared to low achievement definitions of reading disability: Results from the Connecticut Longitudinal Study. *Journal of Learning Disabilities,* 25**,** 639-648.

SHAYWITZ, S. E., ESCOBAR, M. D., SHAYWITZ, B. A., FLETCHER, J. M. & MAKUCH, R. (1992b) Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. *New England Journal of Medicine,* 326**,** 145-50.

SIEGEL, L. S. (1989) IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities,* 22**,** 469-478.

SIEGEL, L. S. (1992) An evaluation of the discrepancy definition of dyslexia. *Journal of Learning Disabilities,* 25**,** 618-629.

SILVA, P. A., MCGEE, R. & WILLIAMS, S. (1985) Some characteristics of nine-year-old boys with general reading backwardness or specific reading retardation. *Journal of Child Psychology and Psychiatry,* 20**,** 407-421.

SNOWLING, M. (2005) Specific learning difficulties. *Psychiatry,* 4**,** 110-113.

STANOVICH, K. E. (1991) Discrepancy definitions of reading disability: Has intelligence led us astray? *Reading Research Quarterly,* 26**,** 7-29.

STANOVICH, K. E. (1993) A model for studies of reading disability. *Developmental Review,* 13**,** 225-245.

STANOVICH, K. E. (2005) The future of a mistake: Will discrepancy measurement continue to make the learning disabilities field a pseudoscience? *Learning Disability Quarterly,* 28**,** 103-106.

STEEL, S. & FUNNELL, E. (2001) Learning multiplication facts: A study of children taught by discovery methods in England. *Journal of Experimental Child Psychology,* 79**,** 37-55.

STEIN, J., TALCOTT, J. & WALSH, V. (2000) Controversy about the visual magnocellular deficit in developmental dyslexics. *Trends in Cognitive Sciences,* 4**,** 209-211.

STEIN, J. & WALSH, V. (1997) To see but not to read; the magnocellular theory of dyslexia. *Trends in Neuroscience,* 20**,** 147-152.

STEIN, J. F. (2001) The magnocellular theory of developmental dyslexia. *Dyslexia,* 7**,** 12-36.

STEVENSON, J. (1988) Which aspects of reading ability show a 'hump' in their distribution? *Applied Cognitive Psychology,* 2**,** 77-85.

STUEBING, K., K., FLETCHER, J. M., LEDOUX, J. M., LYON, G. R., SHAYWITZ, S. E. & SHAYWITZ, B. A. (2002) Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal,* 39**,** 469-518.

SWANSON, H. L. & JERMAN, O. (2006) Math disabilities: A selective meta-analysis of the literature. *Review of Educational Research,* 76**,** 249-274.

THORNDIKE, R. L. (1963) *The concepts of over-and underachievement,* New York, Teachers College, Columbia University.

TYMMS, P. & ALBONE, S. (2002) Performance Indicators in Primary Schools. IN VISSCHER, A. J. & COE, R. (Eds.) *School Improvement Through Performance Feedback.* Lisse, The Netherlands, Swets and Zeitlinger.

VAN DER WISSEL, A. & ZEGERS, F. E. (1985) Reading retardation revisited. *British Journal of Developmental Psychology,* 3**,** 3-9.

VAUGHN, S. & FUCHS, L. S. (2003) Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research and Practice,* 18**,** 137-146.

VELLUTINO, F. R., SCANLON, D. M. & LYON, G. R. (2000) Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ-achievement discrepancy definition of reading disability. *Journal of Learning Disabilities,* 33**,** 223-238.

VIDYASAGAR, T. R. & PAMMER, K. (2010) Dyslexia: a deficit in visuo-spatial attention, not in phonological processing. *Trends in Cognitive Sciences,* 14**,** 57-63.

VON ASTER, M. G. & SHALEV, R. S. (2007) Number development and develop-mental dyscalculia. *Developmental Medicine and Child Neurology,* 49**,** 868-873.

VUKOVIC, R. K. & SIEGEL, L. S. (2006) The double-deficit hypothesis: A comprehensive analysis of the evidence. *Journal of Learning Disabilities,* 39**,** 25-47.

WAGNER, R. F. (1973) Rudolph Berlin: Originator of the term dyslexia. *Annals of Dyslexia,* 23**,** 57-63.

WEPMAN, J. M., CRUICKSHANK, W. M., DEUTSCH, C. P., MORENCY, A.S. & STROTHER, G. R. (1975) Learning disabilities. IN HOBBS, N. (Ed.) *Issues in the classification of children.* San Francisco, Jossey-Bass.

WILSON, A. J. & DEHAENE, S. (2007) Number Sense and Developmental Dyscalculia. IN COCH, D., DAWSON, G. & FISCHER, K. W. (Eds.) *Human Behavior, Learning, and the Developing Brain: Atypical Development.* New York, Guildford Press.

YULE, W. (1973) Differential prognosis of reading backwardness and specific reading retardation. *British Journal of Educational Psychology,* 43**,** 244-248.

YULE, W., RUTTER, M., BERGER, M. & THOMPSON, J. (1974) Over- and under-achievement in reading: distribution in the general population. *British Journal of Educational Psychology,* 44**,** 1-12.