Testing distributional assumptions in psychometric measurement models with substantive applications in psychology

Molenaar, D.

Publication date
2012

5

Modeling Non-normality due to Differentiation: The Method of Moderated Factor Analysis.

The general differentiation hypothesis states that the strength of the correlations among a set of IQ subtests varies with a given variable. Instances of the general differentiation hypothesis that have been considered in the literature include age and ability differentiation. Traditionally, the differentiation effect is attributed to the varying role of g in the subtest scores of an IQ test with the age or ability variable. We argue that this is only one possible way in which a differentiation effect may arise. We discuss five ways in which differentiation can emerge in the higher-order factor model of intelligence, and demonstrate that these can be tested using moderated factor analysis. Using this method, we study the degree in which the various formal conceptualizations of differentiation can be distinguished statistically. We investigate the age and ability differentiation hypotheses in a real data set using both a traditional method and the method of moderated factor analysis. We conclude that results concerning the traditional method can be misleading. In addition, results concerning moderated factor analysis show no evidence for age differentiation, and weak evidence for ability differentiation.

5.1 Introduction

In intelligence research, differentiation of cognitive abilities concerns the variation in the size of the correlations among cognitive ability tests as a function of a given observed or latent variable (Spearman, 1927). In normal populations, it is well established that the correlations among cognitive ability tests are positive. This has given rise to the hypothesis of a common factor, which is general to all cognitive abilities. This factor is usually called *general intelligence*, and denoted *g* (Spearman, 1904; see also Brody, 1992; Jensen, 1998; Johnson & Bouchard, 2004; Macintosh, 1998). The *g*-factor, as a reliably statistical finding in factor analytic studies, is widely accepted (Gottfredson, 1997), although its substantive interpretation remains a subject of debate (e.g., Bartholomew, Deary, & Lawn, 2009; Borsboom & Dolan, 2006; Jensen,

* This chapter is published as: Molenaar, D., Dolan, C.V., Wicherts, J.M., & van der Maas, H.L.J. (2010). Modeling Differentiation of Cognitive Abilities within the Higher-Order Factor Model using Moderated Factor Analysis. *Intelligence, 38,* 611-624.

1998, chapter 4; Horn & Noll, 1994; van der Maas et al., 2006; Plomin & Spinath, 2002).

The success of the *g*-factor, as a statistical explanation of the positive correlations among cognitive ability tests, is matched by the failure of research into the differentiation hypothesis to produce definite results (see below). Most relevant research has addressed differentiation with respect to general intelligence itself (i.e., ability differentiation; Spearman, 1927) and age (i.e., age differentiation; Garret, 1946; where age is a proxy of psychological development).

The initial focus of differentiation research concerned Spearman's original hypothesis of ability differentiation. The ability differentiation hypothesis has enjoyed a good deal of attention in the intelligence literature. Spearman (1927, p. 218-219) observed that the mean correlation among 12 IQ subtests differed substantially between 78 normal children (.47) and 22 learning disabled children [23] (.78). Spearman hypothesized that the strength of *g* decreases across its range, resulting in higher subtest correlations in the lower regions of the *g*-distribution. This hypothesis is referred to as Spearman's Law of Diminishing Returns. Studies of the ability differentiation hypothesis have produced inconsistent results. Some studies found support for the effect (e.g., Abad, Colom, Juan-Espinosa, & Garcia, 2003; Deary, et al., 1996; Jensen, 2003), some found support for the opposite effect (e.g., Bloom, et al., 1988; Hartmann & Teasdale, 2004), and some studies found no effect at all (e.g., Arden & Plomin, 2007; Facon, 2004; Fagorty & Stankov, 1995; Hartman & Reuter, 2006). A recent exhaustive review showed that results concerning ability differentiation are indeed mixed: 55% of the studies reviewed, which were considered valid, supported ability differentiation, whereas 45% did not (Hartmann, 2006, page 169).

The age differentiation hypothesis states that the correlations among cognitive abilities decrease with age due to a progressively diminishing role of *g* with age (Deary et al., 1996). The research into age differentiation again has produced inconsistent results. Although early studies consistently report the effect (e.g., Asch, 1936; Clark, 1944; Filella, 1960; Garrett, 1946), more recent studies show mixed results. For instance, Horn (1970), Horn & Donaldson, (1980), Wang & Kaufman (1993), Tideman & Gustafsson (2004) reported positive results. However, Bickley, Keith, & Wolfle (1995), Caroll (1993), Juan-Espinosa, Garcia, Colom, & Abad (2000), and Rietveld, Dolan, van Baal, & Boomsma (2003) found no support for age differentiation.

Thus, results concerning age and ability differentiation are largely inconsistent. This inconsistency has been attributed to the diversity of ability measures used in different studies (Austin, Hofer, Deary & Eber, 2000), and to the diversity of ages studied (for instance, van der Maas, et al., 2006; van der Maas & Kan, 2007). Another

[23] Spearman's term is 'defective' children.

possible confound with respect to age differentiation research is age dedifferentiation (e.g., Juan-Espinosa et al., 2002).[24] In addition, inconsistency in the results may be due to a possible interaction between ability and age differentiation. This would render the individual effects hard to detect when studied in isolation (as suggested by Arden & Plomin, 2007 and Facon, 2006). However, Tucker-Drob (2009) found no evidence for such an interaction.

We do not claim that the discussion above is an extensive overview of the literature (see Hartmann, 2006, for a review). Rather it shows that results are quite inconsistent. As an anonymous reviewer pointed out, a meta-analysis can shed light on the question whether this body of literature supports differentiation. However, we think that the inconsistency of the results has two causes, which should be addressed at source. First, the methods used are suboptimal because they are based on ad hoc procedures, especially in the case of ability differentiation. Second, an explicit framework concerning the locus of differentiation effects in the higher-order factor model of intelligence is absent. Manifest differentiation is considered an effect on the size of the subtest correlations. Within the higher-order common factor model, such effects can be brought about in a variety of ways, as we discuss below. The use of various ad hoc methods is perhaps due to the fact that the statistical modeling of differentiation effects is no simple matter compared to, say, fitting linear factor models to cognitive ability test scores. However, we suspect that the reliance on ad hoc methods is in part a consequence of the lack of an explicit framework. To return to the reviewer's point: the success of meta-analysis hinges in part on the quality of the results that enter the analysis. The quality of the results depends on the method used. Ultimately any improvement in statistical modeling will ultimately benefit meta-analyses, as it will improve the quality of the empirical results, i.e., the input for meta-analyses.

Our present aim is to present the method of *moderated factor analysis* – a generalization of a method proposed by Bauer & Hussing (2009) and Purcell (2002) - as a general approach to model differentiation within the higher-order factor model. We assume that we have a possibly fallible measure of the differentiation variable (i.e., age or general intelligence). We discuss the statistical and conceptual advantages of this approach in the light of the statistical methods that have been used in the past. We discuss how differentiation may arise in terms of the moderation of the parameters in the higher-order factor model. In so doing we identify specific statistical hypothesis, which are consistent with differentiation effects. We show how the present approach can be used to model and to test these well defined differentiation hypotheses. As differentiation can be due to distinct sources, which are not mutually exclusive, we

[24] We thank the anonymous reviewer that pointed out this possibility.

establish which sources can be resolved reliably. While the presentation in this Chapter concerns age and ability differentiation, other differentiation hypotheses are amenable to the present framework (e.g., neuroticism differentiation; see Austin, Deary, & Gibson, 1997; Bonaccio & Reeve, 2006).

The outline of this Chapter is as follows. We first indicate why past methods are suboptimal. Then, we explicate how differentiation may arise in both the measurement part and the structural part of the higher-order factor model. We present the method of moderated factor analysis to model and to test the presence of differentiation. Then, we investigate the extent to which the different formal conceptualizations of differentiation can be resolved statistically. We perform both moderated factor analysis and a traditional analysis on a real dataset to investigate age and ability differentiation. We end with a general discussion.

5.1.1 Overview of previous methods

From the perspective of statistical modeling, ability differentiation, where differentiation is viewed as an effect of g, is the most problematic instance of differentiation, because *g* is unobserved and internal to the data. That is, *g* is the differentiation variable, but also the main source of individual differences in the test scores. A common procedure followed in ability differentiation research is to create subgroups that differ in *g*, using the scores on one of the subtests of an intelligence test. These groups are then compared using standard statistics, like the (mean) subtest correlations (e.g., Detterman & Daniel, 1989) and the first principal component of the subtest scores (PC1; e.g., Deary et al., 1996; Jensen, 2003). This approach of creating subgroups using observed subtest scores is questionable because the factor structure as established in the complete population is usually distorted in the created subgroups. This may result in a distortion of the subtest correlations, which is unrelated to differentiation (see Meredith, 1964; Muthén, 1989; Nesselroade & Thompson, 1995). An alternative approach that is to compare existing groups that differ with respect to their mean *g*-score (e.g., Detterman & Daniel, 1989; te Nijenhuis & Hartmann, 2006; Reynolds & Keith, 2007; Spearman, 1927). This is not ideal, as any difference in the correlation structure between the groups may be due to *g*, but also to any other variable on which the groups happen to differ (e.g., lower-order cognitive ability factors, or external variables such as SES).

As selection on the observed subtest scores is problematic, some studies have focused on selecting subgroups on the latent variable (*g*), as this will generally not distort the covariance structure in the subgroups.[25] As the latent variable is unobserved, selection is based on factor scores, i.e., linear combinations of the data that approximate the true factor scores. A problem with many types of factor scores (e.g., Bartlett factor score or regression factor scores; Lawley and Maxwell, 1971) is that they are not structure preserving, i.e., the covariances among the common factors based on the calculated factor scores are not equal to the covariances as estimated in fitting the factor model. Carlstedt (2001) and Reynold & Keith (2007) avoided this problem by using Anderson-Rubin factor scores (Saris, de Pijper, & Mulder, 1978), which are structure preserving (see also, Jöreskog, Sörebom, du Toit, & du Toit, 1999; p. 159). This approach is certainly preferable, but still has some problems. First, the number of subgroups to be created is arbitrary. Creating a small number of groups (e.g., two by median split; see Reynolds & Keith, 2007) will result in a violation of the assumption of a multivariate normal distribution for the data, as the factor distribution (i.e., the *g*-distribution) is non-normal within each subgroup. This problem can perhaps be avoided by creating more subgroups (e.g., 7) as then, one is practically conditioning on the *g*-factor, making the multivariate distribution of the data more normal within each subgroup given that the residuals are normally distributed. However, the conditional variance of *g* will be reduced to a small value, i.e., the *g*-factor will be rendered inconsequential as a source of individual differences within each subgroup. Consequently, hypotheses concerning the possible role of *g* in the differentiation effect are hard to test (see Carlstedt, 2001). In addition, the number of subgroups to be created is likely to affect the power to detect differentiation, if present.

As with ability differentiation, the age differentiation hypotheses is commonly investigated by creating subgroups using age variables. These subgroups are then compared with respect to (mean) subtest correlation (e.g., Juan-Espinosa et al., 2000), PC1 (e.g., Li et al., 2004) and factor structure (e.g., Zelinski & Lewis, 2003). This is feasible, as it is straightforward to condition on age. However, with a large age range, the analysis is impractical as it involves the comparison of a potentially large number of groups. For instance, with an age range of 6-89 year (as in the study by Li et al., 2004), one may end up with a large number of groups. A practical solution might be to categorize the continuous age variable into a relatively small number of subgroups (e.g., 6; see, Li et al., 2004). However, categorization of a continuous variable is not recommended generally (e.g., Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delany, 1993) as it lowers the information concerning individual differences (i.e., in age), affecting the power to detect a possible effect of that variable (i.e., a differentiation effect). In addition, to investigate the age differentiation hypothesis, order restrictions need to be tested (for instance $\text{var}(g)_{\text{age group1}} < \text{var}(g)_{\text{age}}$

[25] As long as the regressions of the subtest scores on the common factors and *g* are linear and homoscedastic (see Meredith, 1964).

$_{group2}$ < …, etc.) which is possible, but not simple. Finally, categorizing age is not recommended as cut-off points are arbitrary, which may complicate the comparison of results across studies, for instance in meta-analysis (Royston, Altman, & Sauerbrei, 2006).

Thus, in differentiation research the creation of subgroups that differ with respect to the differentiation variable is problematic. An additional problem involves the ad hoc nature of the statistical methods that are used. The majority of the studies into differentiation focused on the comparison of correlations and principal components. This procedure lacks specificity. For instance, when a significant difference between the subgroups is found on PC1, it is unclear whether this difference can be interpreted in terms of a differentiation effect. The difference may indeed be due to a diminishing role of *g*, but it may also be due to heteroscedasticity of the residuals in the factor model. In addition, as differentiation will result in non-normality of the observed data (see Molenaar, Dolan, & Verhelst, 2010), non-normality due to other reasons (e.g., floor and ceiling effects, restriction of range) may be misinterpreted as differentiation effects. Thus, differences between the subgroups on the first-principal component may be due to other sources of non-normality than differentiation. This emphasizes the importance of casting the differentiation hypotheses in terms of definite effects on parameters in an explicit model. Thus, ideally, the analysis of differentiation should include explicit model fitting, competing model comparison, and goodness of fit testing. In this sense, we follow Tucker-Drob (2009), Hessen & Dolan (2009), and Molenaar et al. (2010), who formulated explicit effects of (ability) differentiation on the parameters of the factor structure of intelligence.

5.1.2 Locus of differentiation and specificity

Studies of ability differentiation have focused exclusively on the increasing or decreasing role of *g* in IQ subtest correlations. The purported role of *g* requires a precise formulation in terms of model parameters. For instance, should we consider changes in the variance of *g*, or changes in the *g*-loadings? To our knowledge, Reynolds & Keith (2007) were the first to indicate that increases or decreases in subtest correlations may have other causes, which are not related to *g*, such as differences in residual variances across *g*. Thus in developing methods to model differentiation, it is important to conceptualize clearly where in the model differentiation arises (i.e., which parameters account for the differentiation effect). Recently, Hessen & Dolan (2009) proposed a heteroscedastic single factor model to study ability differentiation. In their model, the higher subtest correlations at lower levels of *g* are attributable to relatively smaller residual variance of the observed indicators. Tucker-Drob (2009) proposed a one-factor model with non-linear *g*-loadings to test ability differentiation (where the *g*-

loadings are quadratic) and age differentiation (where the *g*-loadings depended on age). In addition, Molenaar et al. (2010), presented a single common factor model in which ability differentiation may be due to heteroscedastic residuals (see also Hessen and Dolan, 2009), skewness of the latent distribution of *g*, and variations in the magnitude of the factor loadings.

This model-based work is not subject to the criticism discussed above. However, these developments concern the one-factor model only (i.e., Hessen & Dolan, 2009; Molenaar et al., 2010; Tucker-Drob, 2009), while multivariate intelligence tests (e.g., the K-ABC-II, Kaufman & Kaufman, 2004; the WISC-III, Wechsler, 1991; and the WAIS-III, Wechsler, 1997) have a multidimensional higher-order structure (see Carroll, 1993; Gignac, 2008; Gustaffson, 1984; Jensen, 1998; Johnson, & Bouchard, 2004). Reynolds, Keith, and Beretvas (2010) proposed factor mixture models to test for ability differentiation in the multidimensional second-order factor model. This approach is certainly promising as a test on ability differentiation. However, using this method, it is not straightforward to test other differentiation hypotheses (e.g., age or neuroticism differentiation).

As differentiation can be conceptualize in a number of ways, the important question arises which conceptualizations of differentiation can be distinguished *statistically* in realistic sample sizes and interpreted *meaningfully*. To our knowledge the only study that addressed these issues within ability differentiation found that an effect on the distribution of *g* is difficult to distinguish statistically from an effect on the *g*-loadings. However, heteroscedasticity is well separable from these two effects (Molenaar, et al., 2010).

5.2 The Higher-Order Common Factor Model and the Differentiation Hypothesis

In factor modeling of multivariate intelligence tests, *g* is modeled as a factor that is common to all IQ subtests. There are two general approaches: the higher-order factor model and the (nested) bi-factor model. We focus on the simplest version of the former, in which *g* is specified as a second-order factor as this is the more widely accepted representation (Gustafsson & Balke, 1993; Jensen, 1998, p.78; Johnson & Bouchard, 2004; Carroll, 1993). More importantly, this model retains the distinction between the measurement (psychometric) model, in which observed subtest scores are related to first-order factors, and what we view as the structural model, in which the (linear) relations among the latent variables (first-order factors and *g*) are specified. This distinction is important as differentiation may in principle originate in either part of the model. However, given the development of moderated factor analysis within the second-order model, the application to the bi-factor model is straightforward, as is its application in third or higher order factor models (e.g., see Carroll, 1993; Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004).

As mentioned, the second-order factor model comprises two parts, a measurement model and a structural equation model. Each part is characterized by its own parameters. A graphical representation of the second-order factor model is depicted in Figure 5.1. In the measurement model the (co)variances among the *p* observed subtest scores are modeled by regressing the subtest scores (indicators) on the *q* first-order common factors.

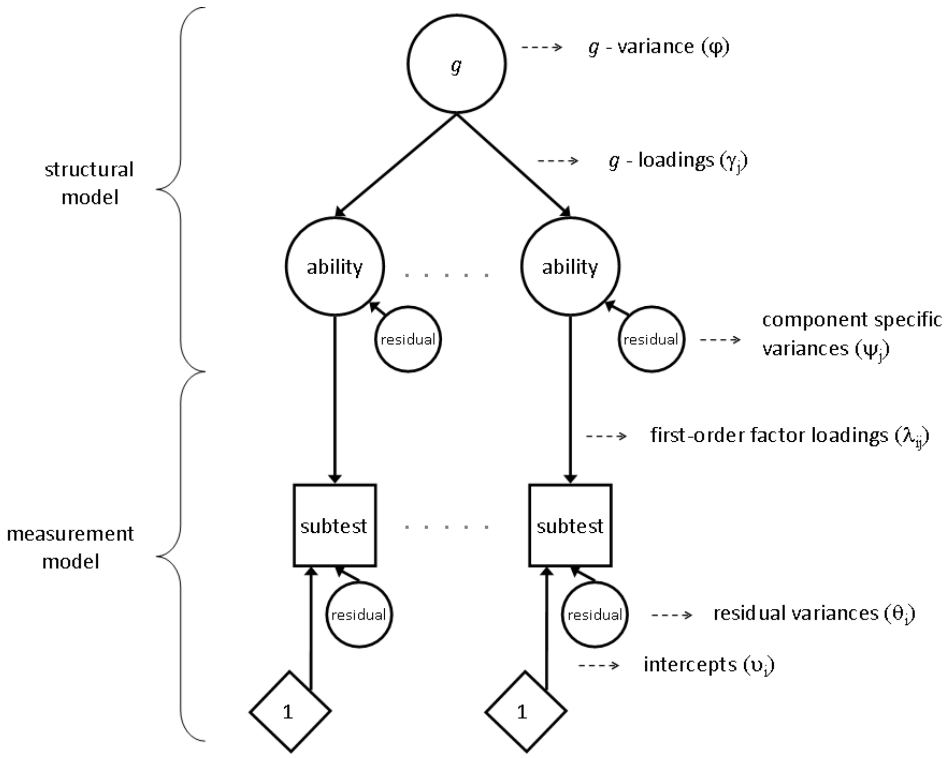

Figure 5.1. The second-order factor model of intelligence. Squares represent observed test scores, circles are unobserved variables, and diamonds are unit vectors (which serve to model the means).

The expected p x p covariance matrix $\mathbf{\Sigma}$ and p dimensional mean vector $\boldsymbol{\mu}$ equal

$$\mathbf{\Sigma} = \mathbf{\Lambda}\, \mathbf{\Omega}\, \mathbf{\Lambda}^{T} + \mathbf{\Theta} \tag{1}$$

and

$$\boldsymbol{\mu} = \boldsymbol{\nu},$$

respectively. The p dimensional vector $\boldsymbol{\nu}$, with elements denoted ν_i, simply contains the means of the subtest scores. There is a one-to-one relation between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, because the model for the means is saturated (all factor means are fixed at zero).[26] The p x p covariance matrix $\boldsymbol{\Theta}$ contains the residual variances of the p indicators. Ideally this matrix is diagonal, although off diagonals (covariances among residuals) can be estimated. As we assume $\boldsymbol{\Theta}$ to be diagonal, and denote the i-th diagonal element of $\boldsymbol{\Theta}$ as θ_i. The p x q matrix $\boldsymbol{\Lambda}$ contains the first-order factor loading of the i-th subtest on the j-th common factor. Let λ_{ij} denote an element in $\boldsymbol{\Lambda}$. We assume that a sufficient number of factor loadings can be fixed to zero on theoretical grounds, as in standard confirmatory factor analysis. The q x q matrix $\boldsymbol{\Omega}$ is the covariance matrix of the first-order common factors. In the structural model, this covariance matrix is modeled by regressing the first-order common factors on g. The covariance matrix of the first-order common factors is

$$\boldsymbol{\Omega} = \boldsymbol{\Gamma}\, \boldsymbol{\varphi}\, \boldsymbol{\Gamma}^{T} + \boldsymbol{\Psi}, \tag{2}$$

where the q x q covariance matrix $\boldsymbol{\Psi}$ contains the first-order factor specific variances of the q common factors. We assume that this matrix is diagonal, and denote the j-th diagonal element of $\boldsymbol{\Psi}$ as ψ_j. The q dimensional vector $\boldsymbol{\Gamma}$ contains the second-order factor loading of the common factors on the second-order factor, g. Let γ_j, denote the j-th second-order factor loading.[27] Finally, $\boldsymbol{\varphi}$ is a 1 x 1 matrix containing the variance of g.

The general differentiation hypothesis states that the correlations among the IQ subtest scores decreases (increases) systematically as a function of the differentiation variable. For instance in age differentiation, g is supposed to become less influential with age, resulting in a decrease in subtest correlations with age. To model differentiation explicitly, we consider the ways in which differentiation can arise in the second-order factor model. In doing so, we want to consider all possible causes of changes in subtest correlations (i.e., not only those related to g). In the second-order factor model, the correlation between two subtests, say y_k and y_l, is given by

[26] In the single group model, we estimate as many intercepts as there are subtests, and constrain all common factors to have zero means. In multi-group models, one can add latent means to investigate measurement invariance (Sörbom, 1974; Meredith, 1993)

[27] This is the standard term, but regression coefficients may be a more appropriate designation, given that we view the higher order part of the second-order factor model as a structual model (relating latent variables to other latent variables), and not as a measurement model (relating observed variables to latent variables).

$$cor(y_i, y_l) = \frac{\lambda_{ij} \times \lambda_{lm} \times \left[\gamma_j \times \gamma_m \times \varphi + I\left(j = m \right) \psi_j \right]}{\sqrt{\left[\lambda_{ij}{}^2 (\gamma_j{}^2 \times \varphi + \psi_j) + \theta_i \right] \times \left[\lambda_{lm}{}^2 (\gamma_m{}^2 \times \varphi + \psi_m) + \theta_l \right]}}, \quad (3)$$

where y_i is an indicator of the *j*-th first-order factor and y_l is an indicator of the *m*-th first-order factor, and *I*(.) is the indicator function which equals 1 if its argument is true and 0 if its argument is false.[28]

Given Equation 3, we note that a decrease (increase) in the subtest correlation of y_i, which loads on the *j*-th first-order factor with an arbitrary other subtest, y_l which loads on the *m*-th first-order factor, can be caused by (see also Figure 5.1):

1) The residual variances (θ_i) increase (decrease) systematically across the differentiation variable.

2) The first-order factor loadings (λ_{ij}) decrease (increase) systematically across the differentiation variable.

3a) For $j = m$, the first-order factor specific variances (ψ_j) decrease (increase) systematically across the differentiation variable.

3b) For $j \neq m$, the first-order factor specific variances (ψ_j) increase (decrease) systematically across the differentiation variable.

4) The second-order factor loadings (γ_j) decrease (increase) systematically across the differentiation variable.

5) The variance of *g* (φ) decreases (increases) systematically across the differentiation variable.

These possibilities are not mutually exclusive. Possibilities 4 and 5 are consistent with the usual notion that decreasing IQ subtest correlations as a function of the differentiation variable are due to the decreasing strength of *g* as a function of that variable. However, possibilities 1, 2, and 3 are additional scenarios in which IQ subtest correlations decrease, but where the effect is not attributable to *g*.

Previous model-based approaches explicitly addressed 3a (Hessen & Dolan, 2009; Molenaar et al, 2010), 4 (Molenaar et al.; Tucker-Drob, 2009), and 5 (Molenaar et al.), as possible models for (ability) differentiation. However, as mentioned these studies were all limited to the one-factor model, which is too simple for multivariate IQ data. We propose moderated factor analysis as a general approach to investigate differentiation with respect to any differentiation variable of interest in any (higher-order) common factor model. This approach requires that the moderator is measured.

[28] Thus, in Equation 3, $I(j = m)$ equals 0 when two subtest are considered that load on different common factors (because then $j \neq m$), and $I(j = m)$ equals 1 when the two subtests load on the same common factor (because then $j = m$).

Note that the methods of Tucker-Drob, Hessen & Dolan, and Molenaar et al., while limited to the one-factor model, do not require a measured moderator.

5.3 Moderated Factor Analysis

Bauer & Hussong (2009) showed how one can tests whether the parameters in the one factor model depend on a continuous measured variable (a moderator, or in the present case, a differentiation variable). Purcell (2002) showed how such a model can be fitted in the freely available Mx program (Neale, Boker, Xie, & Maes, 2006). We extend the approach of Bauer and Hussong (2009) and Purcell (2002) to model differentiation by extending this approach to the second-order common factor model. We refer to the resulting analysis as *moderated factor analysis.*[29] Moderated factor analysis constitutes a systematic approach that can be used to investigate any form of differentiation, and can handle, in principle, combined sources of differentiation (e.g., age and ability differentiation). The method has the advantages 1) that it avoids arbitrary categorization; 2) that it models differentiation effects in explicit loci within the second-order factor model; 3) that it enables goodness of fit tests and likelihood ratio tests on the various differentiation effects. As mentioned, we assume that we have at our disposal a (possibly fallible) measurement of the moderation variable.

In the traditional second-order common factor model given by Equations 1 and 2 (see also Figure 5.1), all parameters are considered fixed across individuals. This assumption is relaxed in heterogeneous factor analysis (Ansari, Jedidi & Dube, 2002; see also Kelderman & Molenaar, 2007), where the parameters are random over individuals. In differentiation research, however, we expect a systematic relationship between the parameters and the given differentiation variable or moderator. We standardize the moderator as this is recommended when fitting models incorporating moderation effects (Jaccard, Turrisi, & Wan, 1990). Following Bauer & Hussong (2009) and Purcell (2002), we model the relationship between the parameters of the common factor model and the moderator using linear functions for the intercepts and loadings (i.e,. the first-order factor loadings and the second-order factor loadings).[30] For the variance parameters (i.e., the subtest residual variances, the first-order factor specific variances, and the *g*-variance), we use exponential functions (see Bauer & Hussong, 2009 and Hessen & Dolan, 2009).

[29] Bauer and Hussong (2009) use the term moderated non-linear factor analysis to include e.g., IRT models as well. As we are focusing on the linear factor model only, we use the term moderated factor analysis.

[30] Moderation of the intercepts should be included in the model as pointed out by (Purcell, 2002). This effect accounts for the correlation between the moderator and the common factor(s). We note that moderation of the intercepts does not cause any differentiation effect as it is not part of the covariance model see Equation 3.

5.3.1 Formal model

In moderated factor analysis, we model the relationship between the parameters from the common factor model and *M*, as explained next. In the following, we assume *M* to be a standardized variable as this is generally recommended when fitting models incorporating moderation effects (Jaccard, Turrisi, & Wan, 1990). If the reader is not interested in the technical details, one can safely skip to section: *5.3.2 conceptual summary*.

5.3.1.1 The main effect of the moderator

When a moderation (interaction) effect between two variables is introduced in a statistical model, the main effect of both variables should be incorporated in the model as well (Nelder, 1994; Purcell, 2002). As Purcell (2002, p. 563) pointed out, neglecting the main effect of the moderator, *M*, on the data, y_i, will result in spurious moderation effects. The main effect of the moderator accounts for a correlation between *M* and y_i, and thus for a correlation between *M* and the factors. By incorporating the main effect of *M*, the variance accounted for by *M* is partialled out of y_i. The remaining variance is then subjected to the test for moderation. If the variance of *M* is not regressed out, this variance may be incorrectly attributed to moderation (see Nelder, 1994). We therefore include the main effect of the moderator by modeling the intercepts conditional on the moderator, $\upsilon_i|M$ by

$$\upsilon_i \mid M = \upsilon_{i0} + \upsilon_{i1}M + \upsilon_{i2}M^2 + \ldots + \upsilon_{ir}M^r = \sum_{s=0}^{r} \upsilon_{is}M^s \quad (4)$$

The conditional intercepts are thus a polynomial function of the moderator. Now, υ_{is} for $s = 1, \ldots, r$ are the main effects of the moderator. The number of higher-order terms that have to be included in the polynomial (i.e., the choice for *r*) depends on the number of higher-order terms that are included in the polynomials of the other functions which will be introduced below (i.e., in Equations 6, 8, 10, 12, and/or 14). For moderation of the intercepts, *r* should be chosen to equal the largest choice for *r* in the other polynomial functions. In the simplest case, i.e., a linear effect of the moderator, $r = 1$, Equation 4 simplifies to

$$\upsilon_i \mid M = \upsilon_{i0} + \upsilon_{i1}M \,, \quad (5)$$

see Figure 5.2.

5.3.1.2 The First- and Second-Order Loadings

The first-order factor loadings are modeled as a function of the moderator variable, *M*, as follows:

$$\lambda_{ij} \mid M = \lambda_{ij0} + \lambda_{ij1}M + \lambda_{ij2}M^2 + \ldots + \lambda_{ijr}M^r = \sum_{s=0}^{r} \lambda_{ijs}M^s \,. \qquad (6)$$

Thus, the factor loadings in the common factor model are an *r*-th degree polynomial function of *M*. In this way, the relation between the factor loadings and the moderator, *M*, could be modeled in any degree of accuracy depending on the choice for *r*. The choice of Equation 6 is based on earlier work by Hessen & Dolan (2009).

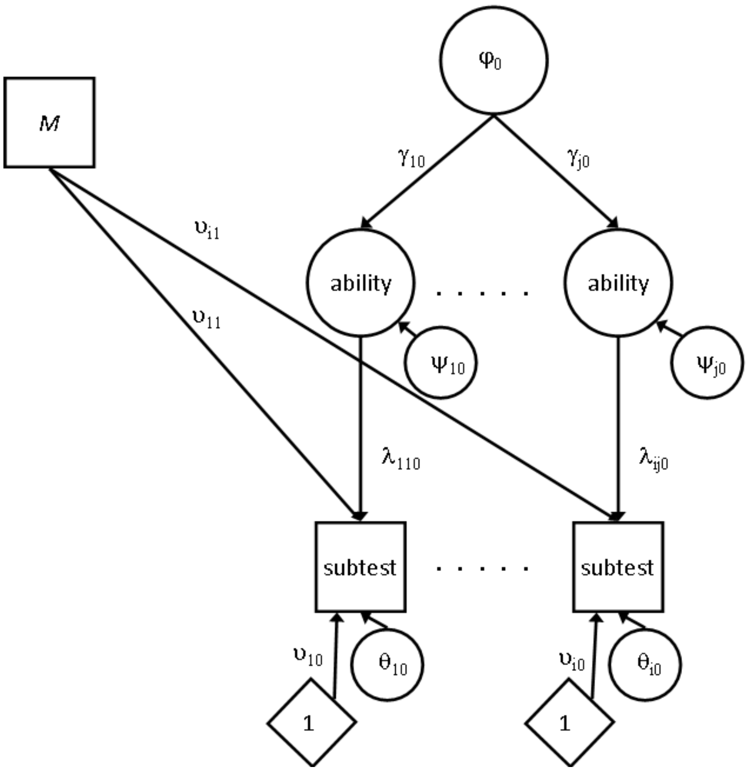

Figure 5.2: Accounting for the main effect of the moderator, M.

In Equation 6, the parameter λ_{ij0} is the *baseline parameter* which accounts for the part of the factor loadings that is independent of the moderator, *M*. The parameters λ_{ij1}, λ_{ij2}, …, λ_{ijr} model the part of the factor loadings that depends on *M*. In doing so, a test on differentiation with respect to the factor loadings would be $\lambda_{ij1} = \lambda_{ij2} = \ldots = \lambda_{ijr} = 0$. With increasing *r* the model becomes increasingly parameterized. For a simple general test of moderation, we adopt the strategy of Hessen & Dolan (2009) and Molenaar et al (2010) and introduce the notion of minimal (linear) moderation, i.e.,

$$\lambda_{ij}|M = \lambda_{ij0} + \lambda_{ij1}M \quad (7)$$

In the Equation 7, λ_{i1} is interpreted as a *moderation parameter*, i.e., if $\lambda_{ij1} = 0$ is rejected (e.g., using a likelihood ratio test, see Buse, 1982, and Hessen & Dolan, 2009), the factor loading of subtest *i* is significantly linearly moderated by *M*.

For the second-order factor loadings we adopt the same approach. That is, the second-order factor loadings (γ_j) are modeled as an *r*-degree polynomial function of the moderator, *M*,

$$\gamma_j \mid M = \gamma_{j0} + \gamma_{j1}M + \gamma_{j2}M^2 + \ldots + \gamma_{jr}M^r = \sum_{s=0}^{r} \gamma_{js}M^s \ . \quad (8)$$

Within this model, the minimal moderation model yields

$$\gamma_j \mid M = \gamma_{j0} + \gamma_{j1}M \ . \quad (9)$$

5.3.1.3 The Residual Variances, First-order Factor Specific Variances, and g- variance

Moderation with respect to the residual variances is modeled in a somewhat different way as these are strictly positive contrary to the factor loadings. We use the natural logarithm of the residual variances (Hessen & Dolan, 2009) resulting in:

$$\ln(\theta_i \mid M) = \theta_{i0} + \theta_{i1}M + \theta_{i2}M^2 + \ldots + \theta_{ir}M^r = \sum_{s=0}^{r} \theta_s M^s \quad (10)$$

The minimal moderation model yields:

$$\ln(\theta_i \mid M) = \theta_{i0} + \theta_{i1}M \quad (11)$$

For the variances specific to the first-order common factors (ψ_j)**,** we adopt the same approach in modeling moderation of the residual variances (Equation 10 and Equation 11). That is, we specify

$$\ln(\psi_j \mid M) = \psi_{j0} + \psi_{j1}M + \psi_{j2}M^2 + ... + \psi_{jr}M^r = \sum_{s=0}^{r} \psi_{js}M^s \tag{12}$$

for the full moderation model, and

$$\ln(\psi_j \mid M) = \psi_{j0} + \psi_{j1}M \tag{13}$$

for the minimal moderation model.

Finally the *g*-variance (φ) may be the source of moderation. As it is a variance parameter we use a same procedure as with the residual variances and the first-order factor specific variances. Thus,

$$\ln(\varphi \mid M) = \varphi_0 + \varphi_1 M + \varphi_2 M^2 + ... + \varphi_r M^r = \sum_{s=0}^{r} \varphi_s M^s \tag{14}$$

for the full moderation model, and

$$\ln(\varphi \mid M) = \varphi_0 + \varphi_1 M \tag{15}$$

for the minimal moderation model.

5.3.1.4 Identification of the Minimal Moderation Model

We focus on testing the differentiation hypothesis using moderated factor analysis. We use the minimal moderation model to test minimal moderation. The minimal moderated common factor model is identified in the same way as the standard common factor model. That is, the means of the *g*-factor, and the specific factors are fixed to 0, one second-order factor loading is fixed to 1, or the variance of *g* is fixed at 1, and for each common factor, one factor loading is fixed at 1.

When moderation is only introduced in the loadings (first- or second-order), *or* in the variances (residual, first-order factor specific, or the *g*-variance), no further identification constraints are necessary. However, for some specific combinations of these effects, additional identification constraints are necessary. First, when moderation is introduced in the second-order factor loadings and in the *g*-variance, one of the moderation parameters of the second-order factor loadings or the *g*-variance should be fixed to zero (i.e., γ_{j1} should be fixed for some *j* or φ_1 should be fixed to zero).[31] Second, when moderation is introduced in the first-order factor loadings and in the first-order factor specific variances, one of the moderation parameters of the first-order factor specific variances should be fixed to zero (i.e., ψ_j should be fixed to zero for some *j*). Or, for each first-order factor specific variances a moderation parameter of the first-order factor loading is fixed to zero (i.e., for each first-order common factor, a λ_{ij1} is fixed for some *i*).

5.3.2 Conceptual summary

In the derivation of the model above, we obtained baseline parameters (i.e., a baseline parameter for each first-order factor loading, a baseline parameter for each second-order factor loading, a baseline parameter for each residual variance, etc), and moderation parameters (i.e., a moderation parameter for each first-order factor loading, a moderation parameter for each second-order factor loading, a moderation parameters for each residual variance, etc.). If a moderation parameter deviates from zero, the corresponding parameter is moderated. For instance, when we take the moderator to be age, and we find that the subtest 'Information' has a significant moderation parameter of 0.5 for the first-order factor loading, we can conclude that the first-order factor loading is moderated by age. That is, the factor loading is increasing with increasing age (because of the positive value of the moderation parameter). In addition, if the *g*-variance is associated with a significant moderation parameter of -0.2, the g-variance is moderated by age. That is, with increasing age, the *g*-variance decreases, because of the negative value of the moderation parameter.

5.4 Simulation study: Specificity of the various effects

To study the practical viability of the model, we investigated the specificity of the various moderation effects by examining the power of the likelihood ratio test to detect various kinds of moderation. The specificity study served the purpose of establishing the degree to which we could distinguish statistically between competing models. The focus is thus on specificity, rather than power per se. As we can consider various causes of moderation, it is important to determine whether moderation effects in one locus (e.g., the subtest residuals) affects tests of moderation relating to other loci (e.g., the first-order factor specific variances). To this end, we used exact data simulation, in which the simulated data fits the model under which it was generated perfectly (i.e., no

[31] Note that fixing φ_1 will eliminate the moderation in the *g*-factor, so generally this is not the best way to identify the model, we mention it for completeness.

sampling fluctuations are introduced in the data; see Bollen & Stine, 1993; Dolan, van der Sluis, & Grasman, 2005; van der Sluis, Dolan, Neale, & Boomsma, 2008). Alternatively one can carry out a standard simulation study. This will produce approximately the same results, but is computationally more demanding.

5.4.1 Procedure

We simulated exact data using the steps described in the appendix. We used 1000 subjects characterized by moderation effects in either 1) the *g*-variance, 2) the second-order (*g*) factor loadings, 3) the first-order factor specific variances, 4) the first-order factor loadings, or 5) the residual variances. In all cases, we used a 3 common factor model with 12 observed indictors. Variables 1 to 4 loaded on the first factor, variables 5 to 9 loaded on the second factor, and variables 10 to 12 loaded on the third factor. We fitted moderated factor models to the generated data in each case. These fitted models incorporated moderation in *either* the *g*-variance, the *g*-loadings, the first-order factor specific variances, the first-order factor loadings, *or* the residual variances. In so doing we can evaluate the power to detect the true effect (e.g., the power to detect moderated *g*-loadings in a model including moderation of the *g*-loadings). More importantly, given our present aim, we can evaluate the specificity of the moderated factor analysis. That is, we want to know whether a given effect (e.g., moderated first-order factor loadings) will be detected as a different effect (e.g., moderated *g*-loadings or moderated residual variances). Power to detect the moderation effect in the data was calculated for each model using the likelihood ratio test (Satorra & Saris, 1985; Saris & Satorra, 1993; see Molenaar, Dolan, & Wicherts, 2009, for an application within IQ research). For the present purpose of investigating specificity, we simulated moderation effects using only a single effect size for each model. These effect sizes are quite arbitrary. The moderation parameters that were chosen for each model are given in Table 5.1 along with the effects sizes in terms of the reliabilities of the subtests (in case of moderation in the measurement model) or the reliabilities of the common factors (in case of moderation in the structural model) for moderator values of -3, 0, and 3.

All other parameters were fixed to certain values (i.e., the baseline parameters of the residual variances and the second-order factor loadings equaled 2; the baseline parameters of the intercepts, the first-order factor loadings, the first-order factor specific variances, and the g-variance equaled 1, and the moderation parameter of the intercepts equaled 2.5). These parameter values resulted in a correlation between the moderator and *g* of .61. Note that in every fitted model, moderation was introduced in the intercepts to account for the main effect of the moderator, as explained above in footnote 8.

Table 5.1
Moderation parameter choices and the corresponding reliabilities of either the subtests or the common factors for M=-3, M=0, and M=3

Model	**Moderation parameter**		**Reliability** *Value of moderator* -3	0	3
Moderated Residual Variances	0.2	tests	.770	.648	.502
Moderated 1st Order Loadings	-0.2	tests	.757	.648	.474
Moderated 1st order residuals	0.3	factors	.908	.800	.619
Moderated *g*-loadings	-0.3	factors	.894	.800	.548
Moderated *g*-variance	-0.3	factors	.908	.800	.619

5.4.2 Results

The diagonals of Table 5.2 contain the power of the different models to detect the corresponding effect in the data. All these coefficients are 1 or near 1. However, these results depend heavily on the particular choices of the parameter values (i.e., effect sizes). Smaller effect sizes, would have resulted in lower power. However, as indicated, we are interested to establish whether specific effects are correctly detected, not in power per se.

The off-diagonals of Table 5.2 contain the power of each model to detect effects that are actually not in the data. Ideally, all these power coefficients are equal to the level of significance (in this case .05) indicating that each model is highly specific, i.e., only those effects that are present are detected.

From the off-diagonal elements of Table 5.2, it appears that it is not possible to separate all moderation effects equally well. For instance, in Table 5.2, if the moderation is limited to the residual variances, then this is detected only using the appropriate likelihood ratio test (i.e., the power to detect the effect is at least .8). So this indicates good specificity of that model to detect this effect. However, moderation of the first-order factor loadings is detected using the appropriate test, but may also be detected, and thus misinterpreted, as moderation of all other parameters, except for the residual variances. Moderation of the first-order factor specific variances is detected using the appropriate test, but may also be detected as moderated first-order factor loadings. Moderation of the second-order factor loadings is again detected using the appropriate test, but may also be detected as moderated first-order factor loadings or as moderated *g*-variance.

Table 5.2
Power to detect different kinds of moderation using moderated factor models. N=1000

Generated data with moderation in:	*Fitted Model* with moderation in: Residual Variances	1st Order Loadings	1st Order Variances	*g*-loadings	*g*-variance
Residual Variances	**1.00** **(12; 388.04)**	0.19 (12; 5.23)	0.61 (12; 2.67)	0.05 (12; 0.02)	0.06 (12; 0.02)
1st Order Loadings	0.14 (12; 7.50)	**1.00** **(12; 262.89)**	**1.00** **(12; 35.02)**	**1.00** **(12; 40.13)**	**1.00** **(12; 40.76)**
1st Order Variances	0.09 (3; 14.52)	**0.80** **(3; 130.90)**	**1.00** **(3; 66.01)**	0.07 (3; 0.52)	0.08 (3; 0.52)
g-loadings	0.05 (3; 0.14)	**0.87** **(3; 151.82)**	0.07 (3; 0.60)	**1.00** **(3; 71.79)**	**1.00** **(3; 71.35)**
g-variance	0.05 (1; 0.11)	**0.87** **(1; 146.50)**	0.07 (1; 0.54)	**1.00** **(1; 70.59)**	**1.00** **(1; 71.64)**

Note. In parentheses are the degrees of freedom and non-centrality parameter (NCP) respectively. Power coefficients above 0.8 are shown in boldface. Alpha was set to equal 0.05.

Finally, moderated *g*-variance may be correctly detected, or may be detected as moderated first-order factor loadings or *g*-loadings. On basis of the results in Table 5.2, we conclude that it is not advisable to fit the moderated factor model with all effects included simultaneously, as some effects are hard to resolve. That is, a model with two unspecific effects together (e.g., moderated *g*-loadings and moderated *g*-variance) could in principle be estimated, however, when a likelihood ratio test is applied (e.g., by dropping the moderation of the *g*-variance) the effect will appear elsewhere in the model (i.e., in the *g*-loadings) and the likelihood function will hardly change. However, the following effects can be studied in combination. Moderation of the residual variances can be combined with any effect (i.e., with moderated *g*-loadings, moderated *g*-variance, first-order factor specific variances, or moderated first-order factor loadings) as these combinations have largely non-overlapping effects. In addition, moderated first-order factor specific variances can be combined with either moderated *g*-loadings or moderated *g*-variance. The decision of which moderation effects to include in the model may be made on theoretical grounds.

5.4.3 Testing differentiation using moderated factor analysis

Using moderated factor analysis it is possible to test the five formal possibilities in which subtest correlations can vary across a differentiation variable. However, we can not resolve all five formal possibilities well. We favor three sources of differentiation effect in view of resolution: moderation of the g-variance, moderation of the first-order factor specific variances, and the residual variances. These sources are statistically resolvable (see Table 5.2) and are well interpretable. Specifically, the first source, moderation of the g-variance, can be interpreted as the classical notion of Spearman (1927) that the strength of *g* diminishes across the differentiation variable. The second source, moderation of the first-order factor specific variances, can be interpreted as heteroscedasticity in the structural model, which suggests that the differentiation effect is not directly due to *g* but due to the first-order factors instead. The third source, moderation of the residual variances, can be interpreted as heteroscedasticity in the measurement model which suggests that the differentiation effect is caused at subtest level.

Thus, we can combine these three sources of differentiation in a single model. If the differentiation hypothesis is true, we expect a systematic pattern of results. For instance, if moderated residual variances are the true cause of the varying correlations across the differentiation variable, the effect should be present in all subtests in the predicted direction (i.e., decreasing or increasing). We may miss an individual effect, but we consider mixed effects (i.e., decreasing *and* increasing residual variances) as inconsistent with the differentiation hypothesis. In case of the second source, moderated first-order factor specific variances, it appears that no systematic effect on all subtest correlations is possible: If the first-order factor specific variances increase with the differentiation variable, the subtest correlations increase for all subtests loading on the same common factors (see possibility 3a above), while the subtest correlations decrease for the subtest loadings on different common factors (see possibility 3b above). Thus, within the multidimensional factor model, possibility 3 can not account for uniformly increasing or decreasing correlations as a function of the differentiation variable. However, it is still a meaningful explanation for a decreasing/increasing strength of the positive manifold across a given variable in the sense that the first-order factor reliabilities increase/decrease across this variable.

5.5 Real data applications

We analyzed data published in Osborne (1980). We selected a dataset consisting of subtest scores of 477 twins on various tests of cognitive abilities. The 477 twins consisted of 328 white twins with a minimum age of 12 and a maximum age of 20, and 149 black twins with a minimum age of 12 and a maximum age of 18. In addition, 247 twin pairs were monozygotic and 230 twin pairs were dizygotic (of which 180 were same sex twins and 50 were opposite sex twins). According to Osborne (1980), twins were drawn from public and private schools in Louisville, Kentucky, and Jefferson County. SES was scored by the occupation (scale 1-7) and education (scale 1-7) of both the father and mother of the twins, resulting in a score between 4 (very high SES) to 28 (very low SES) for each twin. An ANOVA showed out that black and whites differed in mean SES, $F(1,173) = 22.82$, $p < .001$. Other factors (sex, age, and zygosity) were not associated with SES.

We analyzed the raw subtest scores of 12 test batteries Osborne refers to as 'Basic Test Batteries'. These tests are the Calendar Test (CT), the Cube Comparison Test (CC), the Wide Range Vocabulary Test (WV), the Surface Development Test (SD.), the Form Board Test (FB), the Self-Judging Vocabulary Test (SV), the Paper Folding Test (PF), the Object Aperture Test (OA), the Identical Pictures Test (IP), the Newcastle Spatial Test (NS), the Spelling Achievement Test (SA), and the Mazes Test (MT). See Appendix B for a short description of each subtest. Of each twin pair we assigned one member to one subsample (henceforth twin 1) and the other member of the pair to the second subsample (henceforth twin 2). Thus the twin 1 and twin 2 subsamples consisted of 477 subjects of whom 195 and 205 were males, respectively. In addition, data were missing in respectively 26 and 27 subjects on at least one of the 12 subtests. This poses no problem, as Mx is able to handle missing data provided that they are missing (completely) at random (Shafer & Graham, 2002). The means and standard deviations of the scores for Twin 1 and Twin 2 are depicted in Table 5.3.

We performed an exploratory factor analyses in Mplus (Muthén & Muthén, 2007) on the scores of all members of twin 1 (N = 477). We found a three factor structure to fit well and to be interpretable. Judged by the loadings of the subtests we identified the factors as verbal ability, reasoning ability, and spatial ability. Based on this model, we fitted the second-order factor model using confirmatory factor analysis. We included two cross loadings: see Table 5.4 for the standardized parameter estimates. All variance parameters were significant (including the g-variance) except for the variance of the spatial factor (estimate 1.92, s.e. 1.14). This factor model fitted well to the data of the twin 1 members ($\chi^2(49) = 124.63$, RMSEA = 0.057). As we relied on explorative analyses, we cross validated the model using the data of the twin 2 members.[32] In this group, the model fitted as well ($\chi^2(49) = 114.46$, RMSEA = 0.053).

[32] Strictly speaking this is a pseudo cross validation as the data of the Twin 2 members are not independent of the Twin 1 data.

Table 5.3
Means and standard deviations (sd) of the 13 subtests and the age of the Twin 1 and Twin 2 members.

	Twin 1		Twin 2	
	mean	**sd**	**mean**	**sd**
age	15.30	1.55	15.30	1.55
CT	10.65	7.14	10.71	7.07
CC	5.32	8.76	5.57	8.78
WV	4.11	3.56	4.04	3.50
SD	21.55	12.13	21.06	11.30
FB	9.73	6.90	9.59	6.74
SV	32.81	20.29	33.68	20.26
PF	6.19	4.94	6.17	4.70
OA	3.44	6.09	3.08	5.71
IP	52.96	16.97	53.97	16.08
NS	47.42	21.03	48.04	19.99
SA	30.09	15.74	31.12	15.49
MT	26.70	9.32	26.79	9.90

Table 5.4
Standardized parameter estimates for the baseline model in the Twin 1 sample.

Subtest	**Verbal Ability**	**Reasoning Ability**	**Spatial Ability**
WV	.67	-	-
SV	.96	-	-
SA	.80	-	-
CT	.28*	.51	-
CC	-	.69	-
SD	-	.81	-
FB	-	.77	-
PF	-	.80	-
OA	-	.65	-
NS	-	.67*	.28
IP	-	-	.59
MT	-	-	.52
***g*-loadings**	.81	.85	.90
Factor variances	0.35	0.28	0.19

Note. * Two cross loadings were allowed to obtain an acceptably fitting baseline model

5.5.1 Tradition analyses

We first analyzed the data using one of the traditional methods used in the literature. We chose an approach based on principal component analyses because this method has been used most in this context (e.g., Abad et al., 2003; Deary et al., 1996; Jensen, 2003; Juan-Espinosa et al, 2000; Kane & Brand, 2006; te Neijenhuis & Hartmann, 2006). We conducted the test as follows: We median split one of the twelve subtests to obtain a high and low *g*-group. Then, within each group, the Eigenvalue of the first-principal component of the remaining eleven subtests was calculated. Both Eigenvalues were tested on significance using the F-ratio (high divided by low). To see whether the choice of subtest had any effect, we used all subtests in succession to create the high and low ability groups. To test for age differentiation, we followed the same procedure except that we median split on the age variable and that we calculated the first Eigenvalue on all twelve subtests.

Table 5.5.
First principal component in the high and low group based on a median split of each of the subtests and age.

	Twin 1			**Twin 2**		
Variable used for selection	**group**			**group**		
	High	*Low*	*F*	*High*	*Low*	*F*
WV	133.43	85.64	1.56*	155.42	94.97	1.64*
SV	114.50	77.22	1.48*	116.58	69.66	1.67*
SA	123.58	91.12	1.36*	126.21	96.82	1.30
CT	103.81	72.82	1.43*	118.15	67.82	1.74*
CC	122.15	79.77	1.53*	130.23	89.83	1.45*
SD	103.61	66.17	1.57*	123.49	74.55	1.66*
FB	105.70	67.26	1.57*	124.08	67.36	1.84*
PF	97.82	63.14	1.55*	108.46	65.50	1.66*
OA	126.08	85.87	1.47*	137.04	88.93	1.54*
NS	85.47	52.61	1.62*	95.59	57.70	1.66*
IP	138.90	99.18	1.40*	152.96	113.02	1.35
MT	125.29	85.05	1.47*	146.70	105.49	1.39*
age	120.27	108.52	1.11	134.78	127.77	1.05

*Note.**: $p < .01$

Results are in Table 5.5. In case of ability differentiation (i.e., groups are formed on basis of the subtests), we found a significant difference between the Eigenvalue in the high and low group in most of the cases (twin 1: all twelve cases; twin 2; ten cases).

However, the value of the first Eigenvalue is higher in the high ability group than it is the low ability group, indicating that g is a stronger source of individual differences in the high ability group. This effect is taken as evidence against the ability differentiation hypothesis (e.g., Hartmann & Teasdale, 2004). We return to this later. In case of age differentiation (see final row in Table 5.5), we found no significant difference on the first Eigenvalue between the high and low age groups. Thus, the age differentiation hypothesis is not supported by these analyses.

5.5.2 Current procedure

We proceeded by analyzing the data using moderated factor analysis. In Study 1 we investigated age differentiation by taking the age of the subjects as moderator. In Study 2 we investigated ability differentiation hypothesis by taking an external (i.e., to the model) proxy for *g* as moderator. The moderator was standardized, as required, and the subtests were scaled to have approximately equal variances (merely to facilitate computation in Mx). In both studies, we fitted models to the twin 1 data, and we cross validated in the twin 2 data. As explained above, we only tested moderation of the g-variance, the first-order specific variances, and the residual variances. We first fitted a baseline model that included all effects; we then dropped the effects one at the time to test each effect. We retained the effects that showed a significant improvement in the fit of the model. We fixed non-significant parameters to zero. We started by dropping the effect on the *g*-variance, and continued by dropping the effects on the first-order specific variances and the residual variances, respectively. We choose to tests the effects in this order for substantive reasons, i.e., starting by dropping the effect on the g-variance is in our perspective the strongest and most direct test for ability differentiation (in the light of Spearman, 1927), because it concerns the effect on *g* itself. We do not expect that the order in which the effects are tested influenced the main conclusions, as we showed that the effects are statistically well resolvable. We fitted models using Mx (Neale et al, 2006). The documented input files and the data files are available from www.dylanmolenaar.nl.

5.5.3 Study 1: Age differentiation

In each twin group the analyses were limited to the data of the 328 white twin members. The black twins were excluded because we found differences in the factor structure between the black and white twins (specifically, residual variances and first-order factor loadings were found to differ between both groups). These differences can not be a manifestation of age differentiation as all black and white twins are of approximately the same age. Including the data of the black twin members could thus distort the results.

5.5.3.1 Results and conclusions

The model with age moderation of the *g*-variance, the first-order factor specific variances, and the residual variances served as a baseline model. Parameter estimates and standard errors of the moderation parameters are shown in Table 5.6.[33] To test for moderation of the *g*-variance, we dropped the parameter concerning this effect from the baseline model. It appeared that model fit was not affected significantly (Twin 1: χ^2 (1) = 0.755, p = .38, Twin 2: χ^2 (1) = 0.115, p = .73) indicating that the *g*-variance is not moderated by age. We therefore continued with a model with moderation in the first-order factor specific variances and residual variances. We drop the three parameters associated with the age moderation of the first-order factor specific variances. This appeared not to influence the model fit in the twin 1 data (χ^2 (3) = 4.05, p = 0.26), but it did in the twin 2 data (χ^2 (3) = 13.32, p < 0.001). Judged by the standard errors in the twin 2 sample, this effect appeared to be mainly caused by the verbal ability factor. That is, the first-order factor specific variance of the verbal factor in the twin 2 sample is moderated by age (i.e., the factor specific variance is increasing across age, causing the subtests of this factor to get more correlated for the higher ages). However, this effect was not replicated in the twin 1 sample. Note that the spatial factor is associated with a large standard error of the moderation parameter. In the baseline model, the spatial factor was already associated with a large standard error (see above). This may be due to the small number of unique indicators the spatial factor has (i.e., two).

As dropping the moderation of the first-order factor specific variances resulted in a significant likelihood ratio test (at least in the twin 2 sample), we retained this effect, and continued by dropping the moderation in the residual variances. This significantly affected the model fit (twin 1: χ^2 (12) = 24.26, p = 0.02; twin 2: χ^2 (12) = 41.17, p < .001) indicating that at least some residual variances are moderated by age in both samples. Judging by the standard errors of the moderation parameters, we identified the subtests CT, FB, PF, and OA to be moderated significantly by age in the twin 2 sample. The moderation parameter estimate of CT, FB and OA are in the predicted direction (i.e., residual variances increase with age). However the moderation parameter estimate of the PF subtest was in the opposite direction (i.e., the residual variance decreases with age). In the twin 1 sample, none of the univariate effects were significant, notwithstanding the significant multivariate effect.

[33] Standard errors are not part of the default output in Mx because of the optimization procedure used. There are multiple options to determine these, i.e., using the inverse approximate Hessian, using confidence intervals based on the profile likelihood, and using a parametric bootstrap. All standard errors provided in Study 1 and 2 are determined using 100 bootstrap samples. In study 1, one outlier was omitted from these samples because of a parameter estimate 8.5 sd's above the mean. In study 2, three outliers were omitted because of a parameter estimate of respectively 6.8, 6.7, and 7.3 sd's above the mean.

Table 5.6
Moderation parameter estimates (standard errors) in Study 1 for the Twin 1 and Twin 2 members.

		Twin 1	Twin 2
Factor	***g*** **variance**	0.10 (0.12)	0.04 (0.13)
	verbal	0.16 (0.10)	0.42 (0.18)
	reasoning	0.00 (0.21)	0.29 (0.17)
	spatial	0.80 (1.08)	-0.73 (1.14)
Residual Variances	**WV**	0.20 (0.10)	-0.12 (0.09)
	SV	2.03 (1.78)	0.26 (1.26)
	SA	-0.02 (0.10)	-0.01 (0.08)
	CT	0.12 (0.10)	0.21 (0.09)
	CC	0.10 (0.09)	0.07 (0.10)
	SD	0.04 (0.09)	0.02 (0.09)
	FB	0.03 (0.09)	0.30 (0.09)
	PF	-0.02 (0.11)	-0.30 (0.11)
	OA	0.04 (0.10)	0.26 (0.10)
	NS	-0.11 (0.14)	0.01 (0.11)
	IP	-0.01 (0.15)	0.17 (0.14)
	MT	0.17 (0.12)	-0.02 (0.11)

From the above we conclude that, although some moderation effects are present in the twin 2 data, these results do not support the age differentiation hypothesis, as 1) the results are not in line with the age differentiation hypothesis because some moderation was found in the opposite direction for the residual variances, 2) the moderation effects that were found are not replicated in the twin 1 sample, suggesting that the effects may be chance findings. In conclusion, there is little evidence for age differentiation in the Osborne data.

5.5.4 Study 2 Ability Differentiation

To test for ability differentiation, we require a proxy for *g* as *g* itself is unobservable. If the proxy is a reasonable approximation of *g* that is correlated highly with *g*, a predicted decrease across *g* (e.g., the variance of *g* decrease across *g*) will imply a similar, but possibly attenuated, decrease as a function of the proxy. The proxy may be any reasonable approximation of the *g* scores. Here we obtained a proxy for the twin 1 members using the scores on PC1 of the 12 subtests of the twin 2 members. Due to the genetic and environmental similarities between twins, the proxy obtained in the twin 2 members is expected to correlate highly with *g* in the twin 1 members. Nesselroade &

Thompson (1995) used a similar procedure. In cross validating the results we used the PC1 scores of the twin 1 members on the 12 subtests as a proxy in the twin 2 sample.[34]

5.5.4.1 Results and conclusions

We obtained the proxy of *g* for the twin 1 members by taking the PC1 scores of the 12 subtests in the twin 2 sample. Because we wanted to base the PC1 scores on complete records, we had 438 subjects in the twin 1 sample, and 435 subjects in the twin 2 sample. In Table 5.7, correlations between the *g*-proxy and the twelve subtests are depicted, standardized within zygosity and sex. In Table 5.8 correlations between the g-proxy and g are depicted as a function of age corrected for zygosity and sex.

Table 5.7.

Correlations of each subtest with the g-proxy in Study 2 corrected for sex and zygosity.

	WV	**SV**	**SA**	**CT**	**CC**	**SD**	**FB**	**PF**	**OA**	**NS**	**IP**	**MT**
Twin 1	.59	.46	.57	.61	.55	.72	.59	.46	.41	.70	.63	.41
Twin 2	.59	.53	.49	.62	.59	.73	.63	.45	.45	.74	.65	.38

Table 5.8

Correlations (standard errors) between the g-proxy and g as a function of age, corrected for sex and zygosity.

age	**Twin 1**	**Twin 2**
12-13	.84 (0.07)	.85 (0.05)
14	.94 (0.04)	.90 (0.04)
15	.86 (0.05)	.84 (0.05)
16	.93 (0.05)	.90 (0.05)
17	.94 (0.03)	.92 (0.03)

[34] As an anonymous reviewer pointed out, it is in principle possible to split the scores on the proxy and compare the factor structure of the data across these subgroups. This omits the problem of selecting on observed scores as in for instance Deary, et al. (1996). However, the same methodological difficulties remain as discussed above with respect to Reynolds & Keith (2007) and Carlstedt (2001).

In the present analyses, we did not exclude the black twin members. As indicated before, residual variances and first-order factor loadings were found to differ between black and white twin members. This difference can be a manifestation of ability differentiation as blacks and whites differ on the *g* proxy, $t(337.04) = 12.03$, $p<.001$. That is, by the ability differentiation hypothesis, it is predicted that subjects high on the proxy will have lower factor loadings and higher residual variances. Thus in the analysis below we leave the black twin members in the sample as it is possible to model the differences in the factor structure (i.e., in loadings and residuals) between blacks and whites as a function of the proxy.[35]

We followed the same procedure as in Study 1. We fitted a model with moderation in the *g*-variance, the first-order factor specific variance, and the residual variances simultaneously. Parameter estimates and standard errors of the moderation parameters are shown in Table 5.9. To evaluate the statistical significance of the moderation effects, we dropped the effects one at the time and carried out likelihood ratio tests. Dropping the moderation of the *g*-variance did not influenced the model fit (twin 1: χ^2 (1) = 1.18, p = .28; twin 2: χ^2 (1) = 0.27, p = .61) indicating that the *g*-variance is not significantly moderated by the proxy. We thus continued with a model with moderation of the first-order factor specific variances and the residual variances. We dropped the moderation of the factor specific variances. The associated tests were significant (twin 1: χ^2 (3) = 14.43, p = 0.002; twin 2: χ^2 (3) = 44.83, p < .001). Judging by the standard errors from Table 5.9, the misfit was due to the reasoning ability factor in both twin samples. It appears that the factor specific variance of the reasoning factor increases with the proxy. As the first-order specific variances are significantly moderated by the proxy, we retained the parameters relating to these effects. We proceeded by dropping the moderation parameters of the residual variances. The associated tests were significant (twin 1: χ^2 (12) = 93.08, p <.001; twin 2: χ^2 (12) = 84.27, p <.001). Judged by the standard errors we identified the residual variances of CT, CC, SD FB, and OA to be significantly moderated by the proxy in the predicted direction in both samples. Note that all these subtests are in the predicted direction. In addition WV and SV tend to be moderated by the proxy in the Twin 1 sample, and SA in the Twin 2 sample.

[35] Note that we do not focus on a multi-group approach here as this is beyond the scope of the present paper. We thus assume that conditional on the proxy, black and white twin members have the same factor structure. We think this is a reasonable assumption. The application of multi-group models for differentiation and the relation with measurement invariance is an interesting topic of its one.

Table 5.9

Moderation parameter estimates (standard errors) in Study 2 for the Twin 1 and Twin 2 members

		Twin 1	Twin 2
Factor	***g* variance**	0.24 (0.22)	-0.07 (0.18)
	verbal	-0.03 (0.12)	-0.09 (0.10)
	reasoning	0.62 (0.31)	0.82 (0.21)
	spatial	-0.03 (1.86)	-1.48 (1.62)
Residual Variances	**WV**	0.22 (0.08)	0.08 (0.08)
	SV	0.40 (0.16)	-0.04 (0.26)
	SA	-0.04 (0.11)	0.15 (0.08)
	CT	0.30 (0.07)	0.28 (0.09)
	CC	0.20 (0.07)	0.27 (0.07)
	SD	0.40 (0.09)	0.19 (0.08)
	FB	0.30 (0.09)	0.45 (0.07)
	PF	-0.05 (0.08)	-0.06 (0.09)
	OA	0.31 (0.07)	0.26 (0.08)
	NS	-0.01 (0.10)	0.04 (0.10)
	IP	-0.08 (0.08)	-0.06 (0.11)
	MT	-0.18 (2.06)	-0.04 (0.10)

From the above, we conclude that in the present data ability differentiation is limited to the reasoning factor. That is, the factor specific variance and five of the subtest residuals of the reasoning factor increased with the *g* proxy. An increase in the factor specific variance of the reasoning factor across the proxy will cause correlations among the reasoning subtests to increase, and correlations among the reasoning and the other subtests to decrease across the proxy (see Equation 3). The latter is consistent with ability differentiation. The former is however inconsistent with the notion that correlations are uniformly decreasing across *g*.

As we showed using the traditional analyses above, the Eigenvalue of the first principal component was higher for the high *g*-group compared to the low *g*-group. Using the present results of moderated factor analysis, we can explain these findings; we found that some residual variances are increasing for increasing *g*, the first principal component will be larger for high *g*-groups compared to low *g*-groups as there is more variance in the subtest scores. In fact this may support a differentiation effect (if the majority of the subtests shows this effect), but using principal component analysis, one may conclude incorrectly that the opposite effect is present.

5.6 Discussion

We discussed five formal possibilities of differentiation of cognitive abilities within the second-order factor model: residual variances, first-order factor loadings, first-order factor residual variances, second-order (*g*) factor, and *g*-variance. We considered the higher (second) order factor model, because this model distinguishes between the measurement or psychometric model (regression of indicators on substantive first-order factors), and what we view as a structural regression model (regression of first-order factors on *g*).[36] We consider it desirable to maintain this distinction, because we want to be able to test competing differentiation models as moderation of parameters in both parts of the model may given rise to differentiation.

While we distinguish five formal possibilities of differentiation, we showed that there are three sources of differentiation that are statistically resolvable. These concern the *g*-variance, the first-order specific variances, and the residual variances. We argued that these sources can be meaningfully interpreted in terms of respectively the strength of *g*, heteroscedasticity in the measurement model, and heteroscedasticity in the structural model. In this study we choose to incorporate moderation of the g-variance rather than the *g*-loadings as this is the strongest formulation of the differentiation effect in the light of Spearman's traditional hypothesis that *g* gets weaker across a given variable (1927). However, one may also consider moderation of the *g*-loadings as Tucker-Drob (2009) did. This test is more elaborate as it involves investigating differentiation effects on each *g*-loading separately.

In applying the model to data, we found that there was no evidence for age differentiation, which is in line with recent findings of Tucker-Drob (2009). However, the age range of the data we used was restricted to only 12 to 20 year old subjects, it is possible that differentiation is limited to earlier development (i.e., before the age of 12). In the ability differentiation study, we found minor evidence for differentiation. Using a proxy for *g*, we found moderation in the predicted direction for the factor specific variance of the reasoning factor. In addition, we found moderation in the predicted direction for the residual variances of five subtests of the reasoning factor. These results provide some support for the ability differentiation hypothesis, although the effects are limited to the reasoning factor. However, the effect on the factor specific variance of the reasoning factor is difficult to interpret as we found it to be increasing across *g*. This has the predicted effect on the correlations between the reasoning subtests and the

[36] We consider a psychometric or measurement model to be one in which observed indicators are related to a latent variable or common factor. As the second-order factor model relates latent variables (first-order common factors) to another latent variable (the g factor), it is not a measurement model. We refer the reader to Mellenbergh (1994) for an explicit discussion of the linear factor model as a measurement model.

remaining subtests, but it has the reversed effect on the correlations among the reasoning subtests. These results are in line with Hessen & Dolan (2009) and Molenaar et al. (2010), who found that ability differentiation effects, defined in terms of heteroscedastic residuals, were not systematic. However, the results are not in line with Tucker-Drob (2009), who found systematic differentiation of all the *g*-loadings of the first-order factors. However, in his study, subtests scores were summed, which means that possible differentiation effects on the residual variances (as we found in the present study) and first-order factor loadings were not considered. Mainly in the age differentiation application results varied between the two twin samples we used, indicating that findings are chance findings and that no systematic effects underlied the data. We are reluctant to draw definite conclusions on the bases of these illustrative analyses. More extensive applications of moderated factor analysis to a variety of test batteries will provide the basis for such conclusions (or will provide a data base for a meta-analysis, as advocated by a reviewer).

We showed that the outcomes of a traditional analysis using principal component analysis can be misleading in interpreting the results in terms of differentiation. If residual variances increase with increasing *g* or age, the differentiation hypothesis is supported, but a principal component analysis may reveal an opposite effect. We believe that this justifies the work on the development of statistical models for (ability) differentiation, such as presented here, and presented by, for instance, Tucker-Drob (2009) and Reynolds, Keith & Beretvas (2010).

While moderated factor analysis is of potential value in the study of differentiation, we acknowledge some limitations. First, in the case of ability differentiation, one requires an operationalization of *g*, i.e., a proxy that can serve as the moderator. This is no trivial requirement, although studies of the covariance structure of cognitive ability tests often include a wide variety of variables, which will facilitate the derivation of such a proxy (see Johnson and Bouchard, 2004; Naglieri and Jensen, 1987). In addition, we allow that the proxy may be fallible, but the power to detect moderation will suffer if the proxy is poor in terms of reliability or validity.

Appendix A: Exact data simulation

We assumed a standard normal distribution for the moderator. We make this assumption purely to facilitate exact data simulation. In real data applications, no distributional assumptions need to be made concerning the moderator. The moderator was divided into 6 equally spaced bins (bin A: between -3 and -2, bin B: between -2 and -1, bin C: between -1 and 0, bin D: between 0 and 1, bin E: between 1 and 2, bin F: between 2 and 3). In each bin, the middle value was taken as the score on the moderator (i.e., -2.5, -1.5, -.5, .5, 1.5, 2.5, respectively). Then, we used the standard covariance formulae (e.g., Bollen, 1989, p. 35) to calculate the population covariance matrix for the observed data within each bin, i.e., conditional on the chosen values of the moderator. Using this conditional covariance matrix, we simulated exact data within each bin (using the function ‘mvrnorm’ in the ‘MASS’ library of the software package R; R Development Core Team, 2007). The proportion of subjects within each bin was calculated according to the standard normal distribution (i.e., in case of 6 bins: .02, .14, .34, .34, .14, and .02 respectively). So in the present situation, simulation of exact data for a total of 1000 subjects required the computation of a covariance matrix within bin A using the standard covariance formulae and a value for the moderator of -2.5. Once this covariance matrix was obtained, raw data was simulated for .02 * 1000 = 20 subjects. Then this procedure was repeated for bin B, C, D, E, and F. All data within the bins were stacked resulting in a dataset for 1000 subjects. In fitting the true model to this data using raw data likelihood estimation in Mx, we consistently recovered the true parameter estimates, for a range of starting values. In dropping an effect, the value of the likelihood ratio equaled the non-centrality parameter (NCP) of the non-central chi-square distribution. Using this information we calculated the power to detect the effect (see Satorra & Saris, 1985; Saris & Satorra, 1993). Changing the sample size (say from 1000 to 500) only requires an adjustment of the NCP, i.e., you do not need to create new data.

Appendix B: Description of the subtests used in this study

Each subtest is described shortly below. These descriptions are taken from Osborne (1980) and are in some cases slightly adapted. See Osborne for literature references for each subtest.

Calendar Test (CT)
This test contains 50 statements about the days of the week. Subjects are asked to judge whether these statements are true or false, e.g., “If today is Sunday, then tomorrow will be Monday”.

Cube Comparison Test (CC)
In each item two drawings of a cube are displayed. Subjects are asked to judge whether these cubes can be the same cubes or must be different cubes.

Wide Range Vocabulary Test (WV)
In each item, a word is presented and subjects are asked to choose an appropriate synonym from the five options offered. For instance “JOVIAL: 1. refreshing 2. scare 3. thickset 4. wise 5. jolly.”

Surface Development Test (SD)
Subjects are confronted with a piece of paper which should be folded to some kind of object.

Form Board Test (FB)
Five shaded drawings of pieces are presented, some or all of which can be put together to form a figure depicted in outline form. Subjects are asked to identify which of the five shaded pieces should be used.

Self-Judging Vocabulary Test (SV)
80 words are presented. For each word, subjects can choose the meaning of the word from six options. In addition, when subjects think they know the word but are unsure about the six answer categories, they can state the answer in their own words.

Paper Folding Test (PF)
For each item, successive drawings illustrate two or three folds made in a square sheet of paper. A drawing of the folded paper shows where a hole is punched in it. The subject selects one of five drawings to show how the sheet would appear completely unfolded.

Object Aperture Test (OA)
A three dimensional object is shown, followed by outlines of five apertures or openings. The subject is to imagine how the object looks from all directions, then to select from the five apertures the opening through which the solid object would pass directly if the proper side were inserted first.

Identical Pictures Test (IP)
For each item, the examinee is asked to check which of five geometrical figures or pictures is identical to a given figure at the left end of the row.

Newcastle Spatial Test (NS)

This test consists of six subtests ranging in difficulty from simple recognition of regular solids to the more complex problems of surface development.

Spelling Achievement Test (SA)

This test consists of 60 words. Each word is pronounced by the examiner, who then uses it in a sentence, and then pronounces it again. Subjects are asked to write down the proper spelling of the word.

Mazes Test (MT)

Subjects are asked to draw a line from the entrance to the exit of a maze without crossing any line or entering blind alleys.