Differentiation of cognitive abilities in the WAIS-IV at the item level

Molenaar, D.; Kõ, N.; Rózsa, S.; Mészáros, A.

DOI
10.1016/j.intell.2017.10.004
Publication date
2017
Document Version
Final published version
Published in
Intelligence

Intelligence 65 (2017) 48–59

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Differentiation of cognitive abilities in the WAIS-IV at the item level☆

Dylan Molenaar[a,*], Natasa Kő[b], Sandor Rózsa[c], Andrea Mészáros[d]

[a] University of Amsterdam, Department of Psychology, The Netherlands
[b] Eötvös Loránd University, Faculty of Education and Psychology, Budapest, Hungary
[c] Washington University School of Medicine, St. Louis, MO, USA
[d] Eötvös Loránd University, Bárczi Gusztáv Faculty of Education, Budapest, Hungary

ARTICLE INFO

Keywords:
Spearman's law of diminishing returns
Ability differentiation
Age differentiation
Non-linear factor analysis

ABSTRACT

It is known that studying the differentiation of cognitive abilities is associated with many methodological challenges. In the recent years, methods have been developed to address these challenges. However, these methods require that the item scores of an intelligence test are combined into a composite score which may affect the power to detect the differentiation effect or even produce spurious results. Therefore, in this study, an item level approach is presented that can be used to simultaneously test for ability differentiation, age differentiation, and age differentiation-dedifferentiation. The new method is investigated in two small simulation studies, and applied to the standardization data of the Hungarian WAIS-IV. Results indicate that the ability differentiation effect is consistently present in the items of the WAIS-IV while there is no consistent age differentiation and/or age differentiation-dedifferentiation effect.

In 1927, Charles Spearman published one of his seminal books on intelligence. In this book, it was shown that the correlations among various intelligence measures were larger for a sample of children diagnosed with learning difficulties than for a sample of children without such a diagnosis. On the basis of these results, Spearman formulated what is now known as 'Spearman's Law of Diminishing Returns', or 'ability differentiation' which postulates that the general intelligence factor, *g*, explains less individual differences in intelligence for increasing levels of *g*.

Ever since this postulation by Spearman (1927), research has been devoted to study the differentiation effect in more depth. In earlier attempts, subgroups differing in *g* were being compared in their *g*-variance or *g*-loadings. The *g*-subgroups were operationalized by either splitting observed intelligence subtest scores (e.g., Deary et al., 1996; Detterman & Daniel, 1989), factor scores (Carlstedt, 2001; Reynolds & Keith, 2007) or by using existing groups that are known to differ in *g* (e.g., Detterman & Daniel, 1989; Spearman, 1927; te Nijenhuis & Hartmann, 2006). However, it has been argued that the use of *g*-subgroups is suboptimal as – depending on the way these subgroups are created – it will produce confounded results or results that depend on arbitrary decisions concerning the criterion on which the subgroups are formed (Molenaar, Dolan, Wicherts, & Van der Maas, 2010; Murray, Dixon, & Johnson, 2013).

Due to this challenge of creating subgroups, more recent studies have used tests for the ability differentiation effect in which no *g*-subgroups are created. Most notably, Tucker-Drob (2009) and Molenaar, Dolan, and Verhelst (2010) proposed the idea to operationalize the ability differentiation effect as the quadratic effect of *g*. As a result, the regular (linear) effect of *g* can be stronger or weaker for larger values of *g* depending on the direction of the quadratic *g*-effect. It can be shown that this approach is similar to the mixture approach by Reynolds, Keith, and Beretvas (2010) and equivalent to the moderation approach by Molenaar et al. (2010).[1] In addition, the more traditional multigroup approaches discussed above in essence also test for a non-linear effect of g as these more traditional approaches test for an interaction between a proxy for g (i.e., the splitted subtest score, the splitted factor scores variable, or the existing groups variable) and g itself (as operationalized by a factor model within each group). That is, the traditional approaches test for a non-linear effect of g as well as the newer methodology.

As the differentiation effect can thus be seen as a non-linear effect of g, one important challenge to all approaches above remains. This challenge is referred to as 'scaling'. By scaling we refer to the scale of the composite scores as defined by the items of a subtest. By composite scores we refer to any observed measure that is constructed to

☆ The research by Dylan Molenaar was made possible by a grant from the Netherlands Organisation for Scientific Research (NWO VENI-451-15-008). We thank the reviewers for their constructive comments on previous drafts of this paper.

* Corresponding author at: Psychological Methods, Department of Psychology, University of Amsterdam, PO Box 15906, 1001 NK Amsterdam, The Netherlands.
E-mail address: D.Molenaar@uva.nl (D. Molenaar).

[1] Specifically, Bauer (2005) shows the relation between mixture models (similar to those of Reynolds et al., 2010) and non-linear factor models (similar to those of Tucker-Drob, 2009). In addition, if, in the moderation approach by Molenaar et al. (2010), *g* is taken to be the moderator, the model of Molenaar et al. and Tucker-Drob (2009) coincide.

http://dx.doi.org/10.1016/j.intell.2017.10.004
Received 14 June 2017; Received in revised form 17 October 2017; Accepted 17 October 2017
Available online 08 November 2017

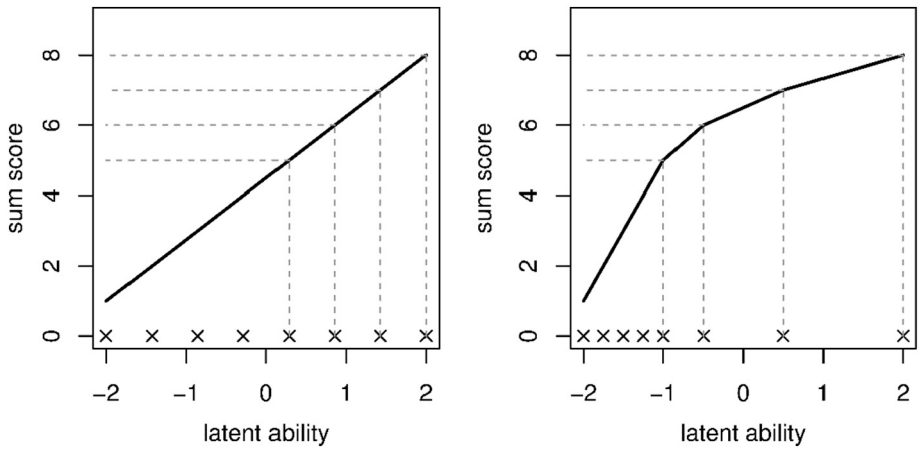

Fig. 1. An illustration of how the scale of the summed item scores is determined by the psychometric properties of the items. Left: the item difficulties (crosses) are located equally spaced across the latent ability scale. As a result, the difference between a sum score of 8 and a sum score of 7 implies the same underlying difference on the latent ability as compared to the difference between a sum score of 6 and a sum score of 5 (grey dashed lines). Right: the item difficulties (crosses) are located disproportionally across the latent ability scale. As a result, the difference between a sum score of 8 and a sum score of 7 implies a *larger* underlying difference on the latent ability as compared to the difference between a sum score of 6 and a sum score of 5 (grey lines).

summarize the performance of the subjects on the items of a subtest. Examples of composite scores include summed item scores (sum scores) and age standardized sum scores. The scale of these composite scores depends on the psychometric properties of the items in the subtest from which the composite scores are calculated. That is, the composite scores from a subtest with many items has a different scale than the composite scores from a subtest with fewer items, and the composite scores from a subtest with more difficult items has a different scale than the composite scores from a subtest with fewer difficult items. These psychometric properties define the scale of the composite scores: it defines how far two subjects are apart on the underlying latent ability scale (e.g., working memory) if the subjects differ only by one point on the composite score. That is, it defines how much two subjects differ in ability if one subject has a composite score of for instance 8 and the other subject has a composite score of 7.

The exact scaling of the composite scores is important as the difference between a composite score of 7 and 8 may not imply the same difference in the underlying latent ability as compared to the difference between a composite score of 5 and 6. The units of the composite score scale are defined by the position of the item difficulties on the latent ability scale. See Fig. 1 for an illustration of this point for the sum score (summed item scores): If the item difficulties are proportionally distributed across the latent ability scale (left plot in Fig. 1), then the difference between a sum score of 7 and a sum score of 8 imply the same difference in latent ability as compared to the difference between a sum score of 5 and a sum score of 6. However, if the item difficulties are disproportionally distributed across the ability scale, the difference between a sum score of 7 and a sum score of 8 does not necessarily imply the same difference in ability as compared to the difference between a sum score of 5 and 6. Note that a ceiling effect is the extreme example of the disproportionality case in Fig. 1 in which there are so many easy items that a substantial part of the subjects obtains the highest possible sum score.

As can be seen in Fig. 1 (right), the scale of the sum scores from a given subtest may already include some non-linearity due to the psychometric properties of the items in that subtest. That is, because in Fig. 1 (right), there are somewhat less difficult items, the unit of the sum scores is larger for higher abilities than for lower abilities. Therefore, in a test on ability differentiation, the sum score from Fig. 1 (right) will spuriously display the ability differentiation effect. Not because ability differentiation is truly in the data, but because the non-linearity in the sum score scale will be wrongfully detected as non-linearity due to ability differentiation. As a result, to be able to test for non-linear effects due to ability differentiation in the sum scores, possible non-linearity due to the psychometric properties of the items should be taken into account (see also Tucker-Drob, 2009). We will demonstrate this later in a simulation study for sum scores and for age standardized sum scores.

The above argumentation holds for all procedures that are based on a transformation of summed item scores, for instance, age standardized sum scores. That is, age standardization procedures can be seen as age-specific transformations of summed item scores. Such age-specific transformations only remove the main effect of age from the data, but these transformations do not remove the possible effects of non-linearity due to the psychometric properties of the items. The only way to account for the properties of the items is to have transformations of the item scores that explicitly take into account the difficulty of the items that are solved correctly.

Some researchers have used Rasch calibrations to account for scaling issues. For instance, researchers may administer many items in a pilot study and only select a subset of the items that have proportional item difficulties. Such calibrations are certainly valuable in test construction, however in the case of testing for ability differentiation, such calibrations may remove non-linearity that is due to the ability differentiation effect. That is, if differentiation is in the data, but this effect is not taken into account during calibration, the item difficulties are biased (due to the differentiation effect which is not taken into account) resulting in potentially wrong corrections of the scale removing the differentiation effect.

Thus the problem is that 1) disproportional item difficulties may produce spurious differentiation effects in composite scores; while 2) it is hard to test whether a composite score is associated with a disproportional scale as the item difficulties are influenced by the presence of ability differentiation. Thus, to be fully confident that in a statistical test on ability differentiation, the effect is a genuine differentiation effect and not a scaling effect, the item difficulties should be explicitly taken into account.

To solve this problem with respect to the scale dependency of the differentiation effect, Tucker-Drob (2009) did not analyze the summed item scores or transformations thereof, but relied on factor score estimates within each of the subtests in the Woodcock-Johnson III standardization data (Woodcock, McGrew, & Mather, 2001).[2] As these factor scores are estimated by taking into account the item difficulties (see e.g., Andersen, 1995), these factor scores are in principle free of the exact scaling due to the item properties. In addition, Molenaar et al.

[2] More specifically Tucker-Drob (2009) used the 'W-scores'. As discussed in McGrew, LaForte, and Schrank (2014), the W-scores are linear transformations (i.e., formula 2.3 on page 46) of the estimates of the ability parameters (B_n) and the difficulty estimates (D_i) from formula 2.1 (for dichotomous data) and from formula 2.2 (for polytomous data). Therefore, these W-scores are a direct transformation of the so-called 'factor scores' (or 'ability estimates') in the Rasch model. Thus, we will refer to the W-scores as factor scores (as the W-scores only differ from the factor scores in their mean and variance, but not in their higher-order moments).

(2010, see also Murray et al., 2013) adjusted the residual variances of the traditional factor model.[3] As a result, possible scale effect could be absorbed in the residual variances such that the tests on differentiation are unaffected in principle.

Although certainly useful, the above solutions are still suboptimal as they work around the more optimal approach of testing for differentiation using the raw items scores. That is, when using factor score estimates to test for differentiation, the standard errors of these factor score estimates are neglected in the actual differentiation tests. By neglecting the standard errors, the factor scores are treated as if they are observed quantities while they are estimations that are subject to estimation error (uncertainty). As a result, the amount of uncertainty in the factor scores is not taken into account in estimating the model parameters. Therefore, the confidence interval of the differentiation parameter (the parameter corresponding to the quadratic effect of g) will be too narrow (as it does not include the uncertainty concerning the factor score estimates) resulting in an increased false positive rate. Furthermore, when using an adjustment of the residual variances to test for differentiation (as is done by Molenaar et al., 2010), a functional form of the adjustment is needed (e.g., a linear effect on the log-variance) for which it is unclear whether this form will successfully capture all scaling effects in the data.

Therefore, in the present paper an item level approach to study ability differentiation is presented based on the non-linear methodology by Tucker-Drob (2009) and Molenaar et al. (2010). In this approach, the psychometric properties of the items in the intelligence subtests are explicitly taken into account without the need of estimating the factor scores first or the need of specifying a functional form for the scaling effects. In addition, the proposed approach will also account for possible effects of age differentiation (Garrett, 1946) and age-differentiation-dedifferentiation (Balinsky, 1941). The outline is as follows: first, the item level approach is presented. Next, in two small simulation studies it is demonstrated 1) that the item level approach does not suffer from spurious differentiation effects while the composite score approach does; and 2) that the item level approach is viable in terms of parameter recovery and the power to separate ability differentiation, age differentiation, and age differentiation-dedifferentiation effects. Next, the item level approach is applied to the item scores of the standardization data of the Hungarian WAIS-IV (Rózsa, Kő, Mészáros, Kuncz, & Mlinkó, 2010).

1. Ability differentiation, age differentiation, and age differentiation-dedifferentiation as nonlinear effects

1.1. Ability differentiation

Within the one-factor model (Molenaar et al., 2010; Tucker-Drob, 2009) and within the second-order factor model (Molenaar, Dolan, & van der Maas, 2012; Murray et al., 2013) the ability differentiation effect can be operationalized as the quadratic effect of *g* (henceforth 'g^2-effect'). This non-linear effect of g is still general in the sense that it can have either a positive effect size or a negative effect size. For differentiation to occur, this effect should be in a certain direction. That is, because the ability differentiation hypothesis predicts a decreasing effect of *g* for higher levels of *g*, the g^2-effect should be negative to be in line with this prediction. Thus, if the g^2-effect is significant, one can only speak of differentiation if the effect is negative (i.e., has a negative regression slope/factor loading).

1.2. Age differentiation and dedifferentiation

As in the approach by Tucker-Drob (2009) age-unstandardized scores are used, the ability differentiation effect cannot be studied independently of the so-called age differentiation and age differentiation-dedifferentiation effects (see e.g., Arden & Plomin, 2007; Facon, 2006; Tucker-Drob, 2009). Age differentiation (Garrett, 1946) refers to the hypotheses that the strength of *g* decreases across the life span while the age differentiation-dedifferentiation hypothesis (Balinsky, 1941) leaves open the possibility that the strength of *g* increases again later in life. As age and *g* are substantially correlated in age-unstandardized data, ability differentiation effects in the data may thus in fact be due to age differentiation effects or age differentiation effects may be due to ability differentiation effects. It is therefore of importance to add the effects of age differentiation and age differentiation-dedifferentiation to our modeling.

With respect to age-differentiation, Tucker-Drob (2009) discussed how this effect can be operationalized by an interaction between age and g ('age × g-effect'; see also Molenaar, Dolan, Wicherts, et al., 2011) on the composite scores. As age-differentiation predicts a decreasing effect of g for increasing levels of age, the age × g-effect should also be negative to be in line with this prediction. Finally, Tucker-Drob (2009) operationalized the age-differentiation-dedifferentiation effect by an interaction between quadratic-age and g ('age^2 × g-effect'). The age-differentiation-dedifferentiation effect predicts a decrease in the effect of g for the earlier ages, but an increase in the effect of g for the later ages, therefore, the age^2 × g-effect should be positive to be in line with this prediction. In the case of age-differentiation-dedifferentiation, it is possible that both an age × g-effect and an age^2 × g-effect are present. In that case, the age^2 × g-effect should be positive (to ensure that the differentiation occurs at the earlier ages and the integration at the later ages instead of the other way around), but the age × g-effect might be positive (i.e., not negative as predicted by the age-differentiation effect) indicating that the differentiation effect for the early ages is weaker than the dedifferentiation effect at the later ages. In addition, the age × g effect might be negative, indicating the opposite (i.e., the differentiation effect being stronger than the integration effect).

1.3. The present item-level model

Ideally, the above methodology is applied to the full factor model of intelligence. However, this would result in a highly complex model with at least three layers of factors, see Fig. 2 for a schematic representation of the full factor model of intelligence subject to differentiation. Because of the complexity of this full differentiation model, researchers have relied on aggregating over the first layer(s) of factors in the factor model by calculating composite scores. For instance, Tucker-Drob (2009) used factor scores for the first layer of factors (denoted 'common factors' in the full factor model for differentiation in Fig. 2) and averaged these for the common factors measuring the same broad factor. The resulting variables constituted a composite score for the broad factors. Next, these composite scores were submitted to the model described above resulting in the model in Fig. 3.

Due to the challenges with the scale of composite scores discussed above, in the present study, we will not rely on composite scores. Instead, we will specify an explicit modeling approach at the item level of the full factor model of intelligence in Fig. 2. Specifically, we will specify a simultaneous model for the item scores and the differentiation effects. As the item scores of intelligence tests items are discrete (e.g., 0, 1, or 0, 1, 2, 3), we cannot use the standard factor model for the items (i.e., the linear factor model for continuous variables as used above). Therefore, we use the discrete factor model (Takane & De Leeuw, 1987; Wirth & Edwards, 2007). Comparable to the standard one-factor model, the discrete one-factor model contains factor loadings which model the relation between a given item and the cognitive ability (common factor). In addition, the discrete one-factor model contains threshold parameters which model the difficulty of a given item (i.e., the degree to which people succeed on that item). An item can have multiple thresholds depending on the number of score categories (i.e., the

[3] Specifically, the residual variances were adjusted to be heteroscedastic instead of homoscedastic.

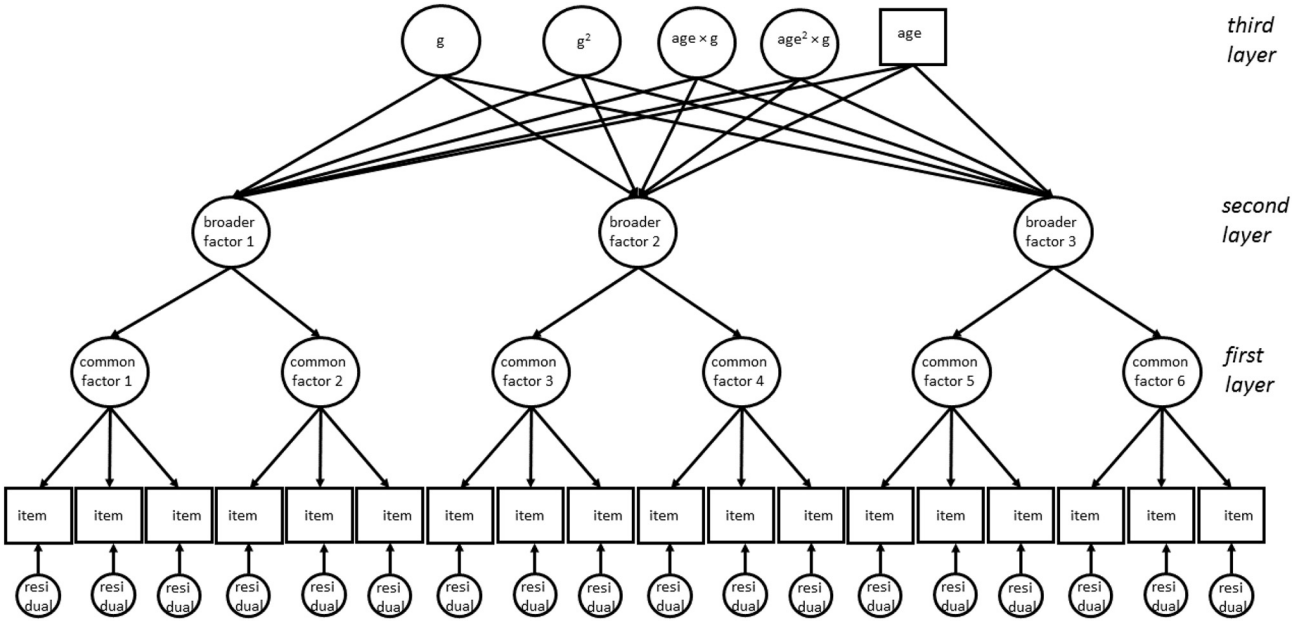

Fig. 2. The full factor model of intelligence subject to ability differentiation, age-differentiation, and age differentiation-dedifferentiation. The number of items, common factors, and broad factors are smaller than in the actual study. Circles represent latent variables, or interactions of latent variables. Single border boxes represent observed variables.

number of thresholds is equal to the number of score categories minus one; e.g., for 2 categories, correct and false, an item has only one threshold, but for 3 categories, the items has 2 thresholds).

Within the discrete one-factor model, we introduce the non-linear effects using the specification as graphically presented in Fig. 4 (see Appendix A for the technical details of the model). We focus on the differentiation effects in the items of the same cognitive ability. That is, as we focus on the first layer in the full factor model of intelligence (Fig. 2) which does not contain g explicitly, we operationalize the ability differentiation effect not as a quadratic effect of g, but as a quadratic effect of cognitive ability (ability^2-effect). Note that this effect thus contains both differentiation effects due to g and differentiation effects due to the cognitive ability specific factors. Similarly, we operationalize the age differentiation effect by an interaction between age and the cognitive ability (age × ability-effect), and we operationalize the age-differentiation-dedifferentiation effects an interaction between age-squared and the cognitive ability (age^2 × ability-effect). These non-linear effects are common to all individual variables (in this case item). That is, if cognitive abilities truly show a differentiation effect, we expect this effect to be evident in all items of that cognitive ability. That is, we test for differentiation effects in the common factors and not in the individual items.

Similarly as in the model for composite scores in Fig. 3, 'ability' (and 'g' in the model for composite scores) does not contain any age variance

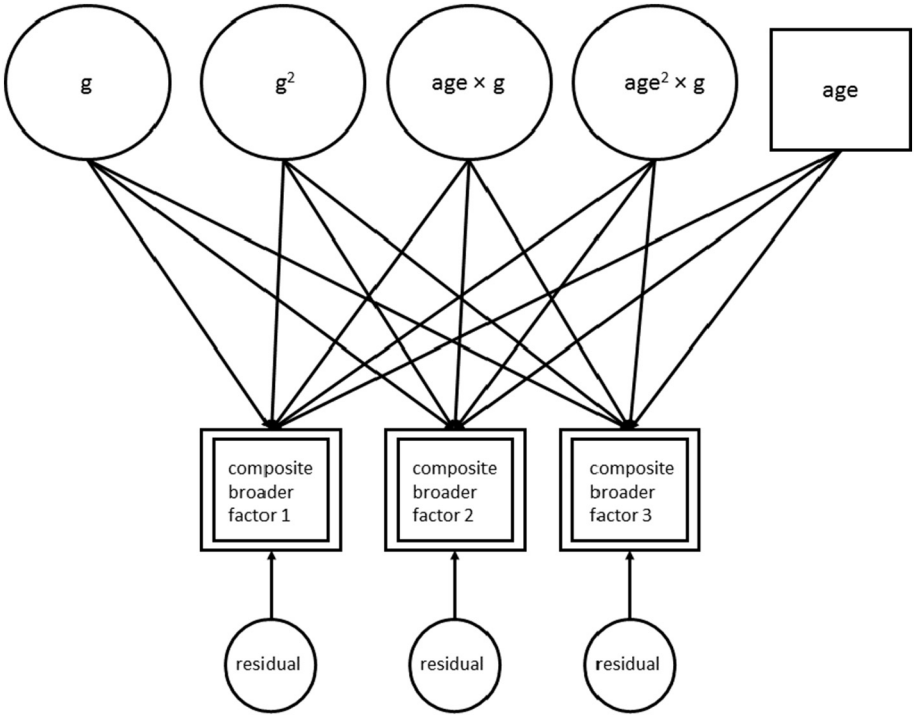

Fig. 3. Schematic representation of the non-linear factor model used by Tucker-Drob (2009) to test for ability and age differentiation in the case of 3 composite broad factor scores (corresponding to the broad factors in Fig. 2). Circles represent latent variables, or interactions of latent variables. Single border boxes represent observed variables, and double border boxes represent scores that are a composite of multiple item scores.

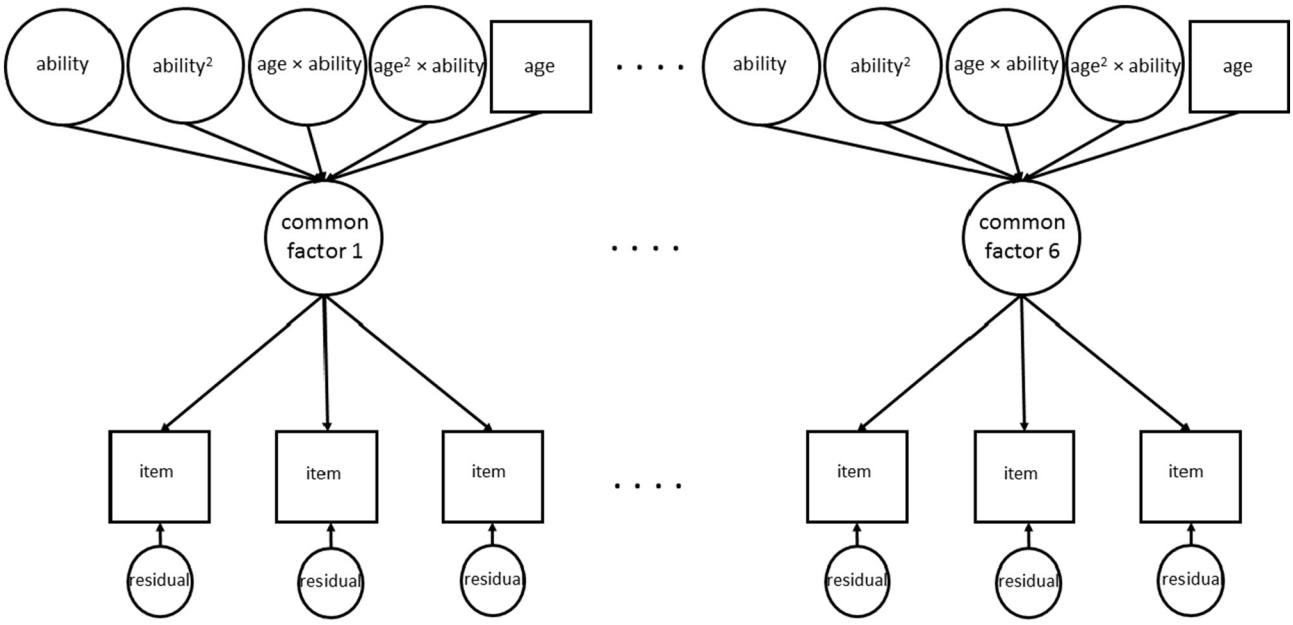

Fig. 4. Schematic representation of the non-linear item level model used in the present study to test for ability and age differentiation in the case of 6 common factors (corresponding to the common factors in Fig. 2). Circles represent latent variables, or interactions of latent variables. Single border boxes represent observed variables, and double border boxes represent scores that are a composite of multiple item scores. The items are explicitly treated as discrete variables, as a result, the residual variances are fixed parameters (but they are depicted in the figure nevertheless). Note that 'ability' represents the common variance of all items after partialling out the effects of age. See Appendix A for the technical details about this model.

as this variance is partialled out by regressing the common factor on age (or by regressing the composite scores on age as in the model for composite scores). Because, in the item level model in Fig. 4, we regress age out at the latent level, we assume that there are no item specific age effects (all age effects run through the common factor). That is, the model assumes measurement invariance across age (the factor loadings and thresholds in the item model are invariant across age) which should be tested prior to fitting the model. We illustrate this in the application to the Hungarian WAIS-IV later.

In the two short simulation studies below, it is illustrated that spurious differentiation effects can arise in the composite score approach (Study A) but not the item-level approach (Study B). In addition, it is shown that the different differentiation effects are well separable in the item-level approach (Study B).

2. Simulation study

2.1. Study A: spurious differentiation effects due to scale properties

In this simulation study, we demonstrate the principle illustrated in Fig. 1. That is, Tucker-Drob (2009) showed visually how spurious differentiation effects can arise in raw sum scores if the item difficulties are disproportionally distributed across the latent ability scale. Here we demonstrate the same, but explicitly considering a statistical test on the differentiation effect to show that the spurious effects are indeed detectable and increase the false alarm rate (Type I error rate). In addition to the spurious effects in the raw sum scores, we also show that age standardized scores suffer from approximately the same increase in false alarm rate.

2.2. Design

Data are generated without differentiation effects for 1000 persons and 4 subtests each consisting of 25 binary items. The correlation between the common factor (cognitive ability) and age was 0.4. For the items, we used factor loadings of 1 for the odd items, and factor loadings of 1.5 for the even items. As we have 25 binary items, we have 25 thresholds (as each item only has one threshold). For these 25 thresholds, we considered two conditions. In the 'disproportional scale condition', we chose 9, 7, 5, 3 and 1 equally spaced values (i.e., in total 25) in the intervals $[-2,-1)$, $[-1,0)$, $[0,1)$, $[1,2)$, and $[2,3]$ respectively.[4] Note that as a result, there are a disproportional number of smaller thresholds, reflecting that there are a disproportional number of easy items. This disproportional distribution of the item thresholds results in different reliabilities across the ability scale, that is, for the lower abilities (below the mean ability) reliability (Cronbach's Alpha) equals around 0.64 and for the higher abilities (above the mean ability) the reliability is around 0.55. In the 'proportional scale condition', we chose 25 equally spaced values in the interval $[-2,2]$ such that there is a proportional number of easy items.

In the simulation study, we analyze the raw sum scores (summed item scores) and the age standardized scores. The age standardized scores are obtained by splitting the age variable in 10 equally sized groups (i.e., N = 100 in each group). Within each group, the raw sum scores are standardized using the mean and standard deviation of the scores within that group. As a result, the raw age variable (i.e., the uncategorized age variable) was uncorrelated with the 4 raw sum score variables. For each condition in the simulation study, we simulated 100 datasets. We fit the model in Fig. 2 to the four raw sum score variables and the four age standardized sum score variables using Mplus (Muthén & Muthén, 2010).

2.3. Results

Table 1 contains the proportion of datasets in which the ability^2-effect (ability differentiation), the age × ability-effect (age differentiation), and the age^2 × ability-effect (age differentiation-integration) was significant for a level of significance of 0.05 in the case of a disproportional scale (disproportional number of easy items) and in the case of a proportional scale. As in this simulation study, the data do not contain any differentiation effects, we refer to these proportions as 'false alarm rates'.

[4] $[-2,1)$ means that -2 is in the interval but -1 is not. So in this case (9 equally space values), we use the following numbers (rounded to 1 decimal place): -2.0, -1.9, -1.8, -1.7, -1.6, -1.4, -1.3, -1.2, and -1.1.

Table 1
False alarm rates for the ability^2-effect (ability differentiation), the age × ability-effect (age differentiation), and the age^2 × ability-effect (age differentiation-integration) in the case of a disproportional scale (disproportional number of easy items) and in the case of a proportional scale.

Scale	Measure	Subtest	ability^2	age × ability	age^2 × ability
Disproportional	Raw sum scores	1	0.90	0.07	0.06
		2	0.88	0.09	0.13
		3	0.93	0.09	0.03
		4	0.94	0.08	0.11
	Age standardized sum scores	1	0.89	0.00	0.02
		2	0.88	0.01	0.03
		3	0.91	0.00	0.01
		4	0.94	0.00	0.01
Proportional	Raw sum scores	1	0.09	0.04	0.10
		2	0.03	0.03	0.07
		3	0.06	0.05	0.06
		4	0.05	0.07	0.09
	Age standardized sum scores	1	0.06	0.00	0.00
		2	0.03	0.00	0.01
		3	0.08	0.00	0.02
		4	0.06	0.00	0.00

2.3.1. Disproportional scale

As can be seen in the table, for a disproportional scale, both the raw sum scores and the age standardized sum scores are associated with false alarm rates that are close to the level of significance (0.05) for the age × ability-effect and the age^2 × ability-effect. This indicates that no spurious effects are detected. For the ability differentiation effect however, that is, the effect of ability^2, both the raw sum scores and the age standardized sum scores are associated with very large false alarm rates (between 0.88 and 0.94) indicating that in a large proportion of the datasets ability differentiation effects are detected while they are not in the data.

2.3.2. Proportional scale

In the case of a proportional scale, the raw sum scores are associated with false alarm rates that are close to the level of significance for any of the effects (i.e., ability^2, age × ability, and/or age^2 × ability). This indicates that there are no spurious effects detected in the data. For the age standardized scores, the false alarm rates are even systematically smaller than the level of significance for the age differentiation (age × ability) and the age differentiation-dedifferentiation effect (age^2 × ability). This is due to the age variance being largely removed from the data by the standardization procedure.[5]

3. Study B: viability of the item level model

3.1. Design

A second series of simulations were intended to demonstrate that the item level analysis does not suffer from spurious differentiation effects in the case of a disproportional scale. In addition, the separability of the age and ability differentiation effects is studied. To this end, we simulated item level data according to Fig. 4 or a subtest consisting of 25 items and responses of 1000 persons using the same disproportional scale as in simulation study A (i.e., we used the same values for the threshold parameters as discussed above). Next, differentiation effects are introduced in the data according to a fully crossed design for the ability^2-effect, the age × ability-effect, and the age^2 × ability-effect. Each of these effects were either 0 indicating the absence of this effect, or −0.3 (ability^2-effect and the age × ability-effect) or 0.3 (age^2 × ability-effect) indicating the presence of the corresponding effect. Note that the parameters are chosen in such a way that they are consistent with respectively ability differentiation, age differentiation, and age differentiation-integration. All other parameters were equal to the parameters used in the first simulation study. We again simulated 100 datasets. To each dataset, we fit the model in Fig. 4. All models are fit in Mplus (Muthén & Muthén, 2010). The syntax to fit the model can be found in Appendix B.[6]

3.2. Results

Table 2 contains the proportion of the datasets for which the ability^2-effect (ability differentiation), the age × ability-effect (age differentiation), and the age^2 × ability-effect (age differentiation-integration) were significant at a level of significance of 0.05 in the various conditions of this simulation study. As for all but one of the conditions, the data do contain one or more of the effects, we refer to these proportions as 'detection rates'. As can be seen from the table, if none of the effects are present, the detection rates (or false alarm rates in this specific case) are all close to the level of significance indicating that no spurious effects are detected. Note that we used the same disproportional scale as in simulation study A above for which the sum scores and the age standardized scores were shown to be heavily biased. From the table it can also be seen that all effects are well separable. That is, if an effect is absent (indicted by a '0' in the table) the detection rate approaches the level of significance (0.05) for that effect. In addition, if an effect is in the data (indicated by a '1' in the table), the detection rates are mostly acceptable (0.8 or larger). Some cases however stand out, for instance, if the ability^2-effect is the only effect in the data, the detection rate is only moderate (0.7). In addition, if the ability^2-effect is in the data together with a age^2 × ability-effect, the detection rate is also moderate (0.74). For all other combinations of effects, the detection rates are considered acceptable.

4. Conclusion

It has been shown that testing for differentiation in the raw sum scores and age standardized sum scores can produce spurious effects for ability differentiation, but not for age differentiation or age differentiation-dedifferentiation. In addition, it has been shown that the item level approach does not suffer from such a bias. In addition, in the item level approach the different effects are well separable such that ability differentiation effects can be detected validly even in the presence of

[5] In an additional simulation study (results available on request), we established that if all differentiation effects are in the data, detecting age differentiation and age differentiation-dedifferentiation using age standardized scores is associated with a large decrease in power as compared to the raw sum scores.

[6] The Mplus code in the appendix is given for 15 items. The length of the Mplus code increases rapidly for an increasing number of items, therefore, we also wrote an R-script that can be used to generate the Mplus scripts for a different number of items. This R-script can be found on the website of the first author.

Table 2
Detection rates for the ability^2-effect (ability differentiation), the age × ability-effect (age differentiation), and the age^2 × ability-effect (age differentiation-integration) for the various data conditions.

Data			Hit rates		
ability^2	age × ability	age^2 × ability	ability^2	age × ability	age^2 × ability
0	0	0	0.03	0.04	0.05
0	0	1	0.06	0.05	0.89
0	1	0	0.02	0.97	0.04
0	1	1	0.04	0.91	0.88
1	0	0	0.70	0.02	0.04
1	0	1	0.74	0.03	0.92
1	1	0	0.78	0.92	0.04
1	1	1	0.81	0.93	0.89

age differentiation or age differentiation-dedifferentiation.

5. Application

5.1. Data

In this section, we analyze the item scores of the standardization data of the Hungarian WAIS-IV (Rózsa et al., 2010). The Hungarian WAIS-IV contains 17 subtests which are completed by 1112 persons with ages between 16 and 90. Table 3 provides an overview of the properties of the items within each subtest. First, the table contains the number of items, Cronbach's Alpha, and the number of score categories of the items. In addition, we fit a discrete one-factor model to the item level data of each subtest (using weighted least squares in Mplus) to see whether the item scores are unidimensional. We consulted the RMSEA (good fit: smaller than 0.05; acceptable fit: between 0.05 and 0.08; and poor fit: larger than 0.08), the CFI and the TLI (good fit: larger than 0.97; acceptable fit: between 0.97 and 0.95; and poor fit: smaller than 0.95) to assess the fit of the one-factor model (see Schermelleh-Engel, Moosbrugger, & Müller, 2003). Using these fit indices, it can be concluded that only VO violates uni-dimensionality according to all indices. For three subtests (i.e., SI, DSf, and LN) only one or two of the fit indices indicates a violation of unidimensionality. For the other subtests, the one-factor model fits acceptable or good according to all fit indices. In the interpretation of the differentiation results below, these results should be kept in mind. That is, if the VO subtest (which clearly violates unidimensionality) is associated with results that stand out from the results of the other subtests, this could be due to the violation of unidimensionality.

Table 3
Number of items (n), Cronbach's Alpha, number of categories (C), and the RMSEA, CFI, and TLI fit measures for a discrete one-factor model for each of the subtests of the Hungarian WAIS-IV.

Subtest	n	Alpha	C	RMSEA	CFI	TLI
BD: Block Design[a]	13	0.84	2	0.05	0.99	0.99
SI: Similarities	20	0.84	3	0.06	0.94	0.94
DSf: Digit Span forwards[a,b]	6	0.70	3	0.09	0.99	0.98
DSb: Digit Span backwards	8	0.72	3	0.07	0.99	0.98
DSi: Digit Span inverse	8	0.76	3	0.08	0.99	0.98
MR: Matrix Reasoning	26	0.91	2	0.02	1.00	0.99
VO: Vocabulary	45	0.92	3[c]	0.09	0.73	0.71
AR: Arithmetic[a]	21	0.87	2	0.05	0.95	0.94
SS: Symbol Search[d]	–	–	–	–	–	–
VP: Visual Puzzles	26	0.90	2	0.02	0.99	0.99
IN: Information[a]	29	0.91	2	0.04	0.98	0.97
LN: Letter-Number sequencing[b]	9	0.78	4	0.09	0.97	0.96
FW: Figure Weights	27	0.90	2	0.04	0.98	0.97
CA: Cancelation	22	0.86	3	0.04	0.96	0.96
CM: Comprehension[d]	–	–	–	–	–	–
PC: Picture Completion	24	0.86	2	0.02	0.99	0.99
CO: Coding[d]	–	–	–	–	–	–

[a] The first item is omitted from the analyses as too few participants failed on this item.
[b] The final item is omitted from the analyses as too few participants succeeded on this item.
[c] For VO, the first 3 items only contain 2 categories.
[d] These subtests do not consist of individual items.

As pointed out by an anonymous reviewer, given the age range within the present dataset (16–90), no age differentiation is to be expected in the present application as age differentiation is generally hypothesized to occur earlier in life. However, from our perspective, age dedifferentiation can still be present in the data as dedifferentiation is hypothesized to occur later in life. Therefore, in the present application, we do include the age x ability and age^2 x ability-effects.

5.2. Results

5.2.1. Measurement invariance across age

As the data are heterogeneous with respect to age, we test for measurement invariance (i.e., invariance of the factor loadings and the intercepts) across age to justify analyzing the age effects at the level of the common factor (cognitive ability) instead of having item specific age effects. To this end, we use the following 13 age categories: 16–17 (N = 100); 18–19 (N = 102); 20–24 (N = 100); 25–29 (N = 100); 30–34 (N = 100); 35–44 (N = 100); 45–54 (N = 102); 55–64 (N = 101); 65–69 (N = 100); 70–74 (N = 51); 75–79 (N = 53); 80–84 (N = 50); and 85–90 (N = 53). These categories have been constructed as part of the stratified sampling scheme in the collection of the standardization data. For the ages smaller than 70, approximately 100 subjects were tested in each age group, and for the ages of 70 and higher, approximately 50 subjects were tested in each age group. The age groups are constructed in such a way that they grasp the relevant developmental information (e.g. 16–17 or 18–19 ranges are narrower as the development of cognitive system is increased).

To the data we fit three models:

Model 1: The factor loadings and the intercepts are free across age groups;
Model 2: The factor loadings equal across age groups while allowing for a difference in the variance of the factor;
Model 3: Both the factor loadings and the intercept parameters are equal across age groups while allowing for both a difference in the variance and the mean of the factor.

To decide on which model is the best fitting, we consider the comparative model fit indices AIC and BIC. Results are in Table 4. As can be seen, for all subtests, Model #3 fits best according to the AIC and BIC,

Table 4
Fit indices for the three models considered in the test on measurement invariance across age.

Subtest	Model 1		Model 2		Model 3	
	BIC	AIC	BIC	AIC	BIC	AIC
BD	16,333	14,889	15,730	14,817	**15,145**	**14,774**
SI	26,017	24,122	25,102	23,989	**24,267**	**23,936**
DSf	7491	6408	6970	6308	**6447**	**6207**
DSb	7769	6691	7260	6598	**6774**	**6533**
DSi	8325	7242	7790	7128	**7273**	**7033**
MR	24,802	21,418	23,095	21,210	**21,353**	**20,972**
VO	66,987	61,307	64,190	**60,971**	**62,674**	61,922
AR	15,767	13,425	14,560	13,237	**13,415**	**13,114**
VP	23,225	19,971	21,478	19,663	**19,799**	**19,428**
IN	28,326	24,937	26,644	**24,759**	**25,881**	25,500
LN	10,094	8840	9559	8782	**8978**	**8679**
FW	21,124	17,863	19,380	17,563	**17,708**	**17,341**
CM	38,802	35,829	37,293	35,579	**35,994**	**35,542**
PC	25,783	22,659	24,139	22,394	**22,714**	**22,353**

Note. For each subtest, the best value of the AIC and BIC fit index is in bold face.

Table 5
Parameter estimates standard errors and Z-statistic for the ability^2-effect (ability differentiation) the age × ability-effect (age differentiation) and the age^2 × ability-effect (age differentiation-integration) in the items of the subtests of the Hungarian WAIS-IV.

Subtest	ability^2			age × ability			age^2 × ability		
	Est	SE	Z	Est	SE	Z	Est	SE	Z
BD	− 0.23	0.02	− 9.42	− 0.09	0.06	− 1.66	− 0.06	0.05	− 1.11
SI	− 0.03	0.03	− 0.86	0.12	0.04	2.76	0.07	0.04	1.86
DSf	− 0.19	0.01	− 13.36	0.05	0.03	1.44	0.04	0.08	0.48
DSb	− 0.17	0.01	− 19.33	0.01	0.01	0.80	0.01	0.02	0.73
DSi	− 0.20	0.00	− 66.33	0.00	0.00	0.33	0.01	0.00	2.00
MR	− 0.15	0.03	− 5.63	0.01	0.05	0.16	− 0.12	0.04	− 3.11
VO	− 0.06	0.02	− 2.75	0.09	0.08	1.13	− 0.05	0.05	− 1.04
AR	− 0.16	0.02	− 7.57	0.01	0.03	0.32	− 0.04	0.03	− 1.31
VP	− 0.20	0.03	− 7.69	− 0.05	0.05	− 1.13	− 0.11	0.04	− 3.11
IN	− 0.004	0.04	− 0.11	0.24	0.08	2.87	− 0.06	0.04	− 1.49
LN	− 0.18	0.02	− 7.95	− 0.04	0.04	− 0.83	0.01	0.04	0.12
FW	− 0.19	0.02	− 9.25	− 0.01	0.06	− 0.19	0.03	0.05	0.65
CM	− 0.04	0.02	− 1.81	0.06	0.03	1.65	− 0.02	0.03	− 0.64
PC	− 0.08	0.03	− 2.52	0.12	0.06	2.05	− 0.03	0.05	− 0.67

Note. BD: Block Design SI: Similarities DSf: Digit Span forwards DSb: Digit Span backwards DSi: Digit Span inverse MR: Matrix Reasoning VO: Vocabulary AR: Arithmetic VP: Visual Puzzles IN: Information LN: Letter-Number sequencing FW: Figure Weights CM: Comprehension PC: Picture Completion.

with the exception that for subtests VO and IN Model #2 is identified as the best fitting by the AIC. However, as for these subtests, the BIC indicates still that Model #3 is the better fitting model we also accept Model #3 for these subtests. We thus conclude that for all subtests, the age effects can be modeled at the level of the factor as there are no substantial item specific effects.

5.2.2. Results for the item level differentiation model

Parameter estimates, standard errors, and Z-statistics of the ability^2-effect, the age × ability-effect, and the age^2 × ability-effect are in Table 5 for each subtest. As can be seen for the ability^2-effect, the effect is in the direction predicted according to the ability differentiation hypothesis for all subtests (i.e., a negative effect). Using a conservative two-sided 0.01 level of significance (lower critical bound: Z = − 2.58), 10 out of the 14 subtests are significant.

As the age differentiation effect (i.e., the age × g-interaction) can only be interpreted in the light of the presence of absence of an age^2 × ability-effect, we first consider the results concerning this latter effect. For the age^2 × ability-effect, it can be seen from the table that not all effects are in the predicted direction (i.e., a positive estimate). As indicated by a conservative two-sided 0.01 level of significance (lower critical bound: Z = − 2.58; upper critical bound: Z = 2.58), 2 out of the 14 subtests are significant. Next, as most of the age^2 × ability-effects are not significant we can interpret the age × ability-effect on its own. It can be seen that not all effects are in the predicted direction according to the age differentiation hypothesis (i.e., a negative effect). In addition, using a conservative two-sided 0.01 level of significance (lower critical bound: Z = − 2.58; upper critical bound: Z = 2.58), only 2 out of the 14 subtests are significant. These two subtests are two different subtests than those that showed the age^2 × ability-effect.

6. Discussion

The present undertaking indicated that for all subtests studied, the ability differentiation effect was in the direction as hypothesized by Spearman (1927). For the majority of the subtests studied (10 out of 14) this effect was also statistically significant. For the age differentiation, and age differentiation-dedifferentiation hypothesis no evidence was found. That is, the effects were not in the predicted direction for most of the subtests and the effects were insignificant for the majority of the subtests (2 out of 14 for both effects). However, as already mentioned, given the age range within the present dataset (16–90), no age differentiation was to be expected as age differentiation is generally hypothesized to occur earlier in life. But with respect to age dedifferentiation, the results are interesting as we did not find convincing evidence for the effect while the effect is hypothesized to occur for the older ages covered by the present data. This absence of an age dedifferentiation effect is in line with the null findings by, for instance, Tucker-Drob (2009) in the Woodcock-Johnson III standardization data and Niileksela, Reynolds, and Kaufman (2013) in the WAIS-IV US standardization sample. However, as pointed out in Niileksela et al. the finding of age dedifferentiation in previous research may be due to severe cognitive decline or dementia of which both are exclusion criteria in drawing standardization samples. This may explain why Tucker-Drob, Nileksela et al., and the present study did not find age dedifferentiation.

In general, we thus conclude that we found evidence for ability differentiation at the item level data of the Hungarian WAIS-IV. As we focused on the item level of the full factor model of intelligence, these results indicate that for a given subtest (e.g., Block Design), the ability measured by that subtest (Block Design ability) explains less variance in the item scores for higher levels of that ability. We thus focused on differentiation across the narrow abilities. Note that these narrow abilities also include variance due to higher-order abilities (e.g., perceptual organization) and variance due to g. Thus, our results do not distinguish between ability differentiation across g, across higher-order abilities, or across narrow abilities. This is thus a 'cost' of the item level approach. However, as the item level approach takes the scaling of the intelligence measures explicitly into account, the scaling confound in testing the differentiation effect can be ruled out. Therefore, we can interpret the effect as a genuine ability differentiation effect which we see as a clear benefit of the present approach.

The cause of the ability differentiation effect (i.e., the process that underlies the effect) is still a topic of investigation. In the literature, mechanisms have been proposed that describe how the differentiation effect arises in intelligence test data. In the 'minimal cognitive architecture' mechanism by Anderson (1992), the ability differentiation effect arises due to differences in the algorithms that are implemented in the cognitive system by high *g* subjects, while low *g* subjects adopt one and the same algorithm. Another example is the 'economic investment' metaphor by Brand (1984) in which high g subjects are compared to rich people who can spend their money to many different things, while low g subjects are compared to poor people who can spend their money only to the basics (e.g., food and rent). In behavior genetics, an explanation for the ability differentiation effect is given by differential heritability across *g* (e.g., Brant et al., 2013; Detterman, Thompson, & Plomin, 1990). More recently, two related explanations have been proposed for the ability differentiation effect. First, in

process overlap theory (Kovacs & Conway, 2016) differentiation is assumed to arise because high ability subjects use more specialized processes (processes with less overlap) than the lower ability subjects. As discussed by Detterman, Petersen, and Frey (2016), in a related theory referred to as system theory (Detterman, Petersen, & Frey, 2001), ability differentiation arises because low ability subjects have weaker central elements in their cognitive system which mask the differences in their less central elements (i.e., the differences in the less central elements are not expressed in the observed scores). For high ability subjects with stronger central elements, the difference in the less central elements are more apparent and thus result in more individual specific variability (and thus relatively less common variability) for the high ability subjects. Thus, as the number of explanations for the differentiation effect is increasing, it might be an interesting next step to combine the theories above into a single comprehensive theoretical framework from which the ability differentiation effect can be interpreted.

In contrast to the above substantive explanations for the ability differentiation effect, there is also an alternative explanation based on selection. Specifically, the differentiation effect might occur due to high ability subjects being systematically under selected in the sampling scheme of intelligence test administrations. This causes g or ability to be a stronger source of individual differences at the lower end of the distribution as there is more variability in g or ability as compared the upper range of the distribution. Whether it is a realistic notion that high ability individuals are systematically under selected is an open question, but the above selection process does provide an explanation for the differentiation effect and should therefore be kept in mind in interpreting the effect.

The simulation studies in the present paper, demonstrating that disproportional item difficulties may result in spurious ability differentiation effects in the sum score but not in the item scores, were meant as a proof of principle. Therefore, the simulations were limited in scope. First, the simulation study was limited in that we only considered scales with a disproportional number of easy items. For a scale with a disproportional number of difficulty items, spurious effects can also arise but in the opposite direction, that is, an ability dedifferentiation effect (see also Tucker-Drob, 2009, who illustrates this visually). Thus, scaling issues can affect ability differentiation results in three ways:

1) A scale with a disproportional number of easy items may produce ability differentiation effects that are not in the data (spurious effects).
2) A scale with a disproportional number of difficult items may produce ability differentiation effects in the opposite direction from the theory (spurious effects).
3) A scale with a disproportional number of difficult items may mask ('cancel out') true ability differentiation effects in the data (false negative).

In ability differentiation research it is thus always advisable to use a method to test for differentiation that takes the scale properties into account (i.e., the difficulty of the items).

Second, in the simulation study, we considered only one configuration for the number of easy/difficult items (i.e., we did not manipulate the severity of the scaling issues). As a result, the question arises whether the scaling issues we introduced are realistic and common in practice. In Fig. 5, we plotted the proportion correct for each item within the seven WAIS-IV subtest that adopt a two category item scoring. In addition, we plotted the proportion correct for each item in the simulation study A (which illustrated that scaling issues can result in spurious effects). As can be seen, our setup (i.e., the distribution of the proportions correct) is not more extreme than those observed in the real WAIS-IV data. It can even be argued that our setup is relatively mild as compared to e.g., the Block Design subtest (BD). In addition, question arises how the item level approach will perform if a subtest consists of only easy items. In such a situation, the item level approach is still feasible in the sense that no spurious results will arise. However, the power to detect an effect will be affected (see Molenaar & Dolan, 2014, who demonstrate this for genotype by environment interactions in item level data).

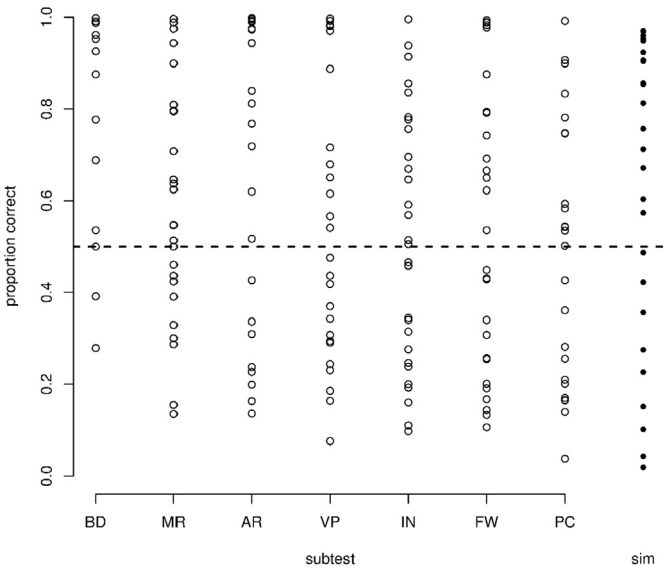

Fig. 5. The proportions correct for the items in the seven WAIS-IV subtests (i.e., BD, MR, AR, VP, IN, FW, and PC) that use two score categories (correct/false) as compared to the proportions correct for the items in an example run of simulation study A (denoted 'sim' in the figure).

In practice, item level data are not always available to the researcher. However, given the danger of spurious results in composite scores with respect to interaction effects in general and differentiation effects in particular, we hope that it will become more and more common practice to share raw (item level) data (see Wicherts & Bakker, 2012). This will enable various robustness checks and sensitivity analyses so that we can hopefully come to well informed and valid conclusions about the differentiation effect and other kinds of interaction effects.

Appendix A

The measurement model

Before we can specify the differentiation effects, we need a measurement model for the item scores to enable a simultaneous modeling approach of the item properties and the differentiation effects. As discussed in the paper, because of the discrete nature of intelligence tests items, we cannot use the linear factor model as a measurement model for the items. Therefore, we use the discrete factor model (Takane & De Leeuw, 1987; Wirth & Edwards, 2007). In the discrete factor model, it is assumed that the discrete item scores have arisen from categorization of an underlying normally distributed variable at specific thresholds. As a result, a linear factor model can be fit to this underlying variable, resulting in intercepts and factor loadings similar as in more traditional factor analyses. As the intercepts and the thresholds are not simultaneously identified, we here fix the intercepts to equal 0. We denote the scores of person p on item i by x_{pi}, and the underlying variable by x_{pj}*. Then, the measurement model is given by

$$x_{pi}^{*} = \lambda_i \times \eta_p + \varepsilon_{pi} \tag{A1}$$

with

$x_{pi} = 1 \text{ if } x_{pi}^{*} < \tau_{i1}$
$x_{pi} = 2 \text{ if } \tau_{i1} < x_{pi}^{*} < \tau_{i2}$
...
$x_{pi} = c \text{ if } \tau_{i(c-1)} < x_{pi}^{*} < \tau_{ic}$
...
$x_{pi} = K \text{ if } x_{pi}^{*} > \tau_{iK}$

where η_p is the common factor, λ_i are the item factor loadings, and τ_{ic} are the item thresholds at which the underlying variable x_{pi}^{*} is categorized. Note that we fixed the intercepts to 0 as these are not simultaneously identified together with the thresholds. Given the idea above, the probability of a response in response category c, $P(x_{pi} = c \mid \eta_p)$, can be determined from the distribution of the underlying variable, x_{pi}^{*}, that is,

$$P(x_{pi} = c \mid \eta_p) = \Phi(\tau_{i(c+1)} - \lambda_i \eta_p) - \Phi(\tau_{ic} - \lambda_i \eta_p) \tag{A2}$$

Introducing the differentiation effects

Next, as discussed in this paper, the ability differentiation, age differentiation, and age-differentiation-dedifferentiation effects are operationalized by focusing on respectively the quadratic effect of ability, the interaction of age and ability, and the interaction of age-squared and ability. That is, in the model for x_{pi} above, the scores of person p on the common factor are modeled by.

$$\eta_p = \nu + \gamma_0 \times \omega_p + \gamma_1 \times \omega_p^{2} + \xi_0 \times age_p + \xi_1 \times age_p \times \omega_p + \xi_2 \times age_p^{2} \times \omega_p \tag{A3}$$

where ω_p can be interpreted as the cognitive ability factor after partialling out the age variable. In addition, ν is an intercept, γ_0 is a factor loading, γ_1 is the quadratic ability effect which captures the ability differentiation effect, ξ_0 is the main effect of age, ξ_1 is the age-ability interaction which captures the age differentiation effect, and ξ_2 is the age^2-ability interaction which captures the age differentiation-dedifferentiation effect.

The full item level differentiation model is then given by Eq. (A2) in which η_{pij} is given by Eq. (A3). Thus the present model is a generalization of the model by Tucker-Drob (2009) in which the Tucker-Drob (2009) parameterization of the differentiation effects are introduced at the latent level and not at the observed variable level. If Eqs. (A2) and (A3) are combined into a single model, two additional identification constraints are needed in addition to the standard identification constraints. That is, in Eq. (A3) ν is not identified as its effect can be captured by the thresholds, and γ_0 is not identified as its effect can be captured by the item factor loadings, λ_i, in A2. Therefore, ν is fixed to 0, and γ_0 is fixed to 1. Note that additionally, the traditional scale and location constraints are necessary. To this end, we fixed the mean and variance of ω_p to 0 and 1 respectively.

Appendix B

The Mplus code below can be used to fit the item level differentiation model to the items of a given subtest. Here, an example is given for 15 items. As can be seen, the code increases rapidly for an increasing number of items, therefore, we wrote an R-code that can be used to generate the Mplus scripts for a different number of items. This R-code can be found on the website of the first author.

```
TITLE: differentiation on item level

DATA:    FILE IS subtest_item_scores.dat;

VARIABLE:          NAMES ARE x1-x15 age age2;

                   MISSING ARE ALL(-999);

                   CATEGORICAL ARE x1-x15;

ANALYSIS:          TYPE = RANDOM;

                   ALGORITHM=INTEGRATION;

MODEL:

 g BY x1-x15*  (l1-l15);

g@1;

g2|g xwith g;

gAge|g xwith age;

gAge2|g xwith age2;

g on age;

g on age2;

x1 on g2 (q1);

x2 on g2 (q2);

x3 on g2 (q3);

x4 on g2 (q4);

x5 on g2 (q5);

x6 on g2 (q6);
```

References

Andersen, E. B. (1995). Polytomous Rasch models and their estimation. In G. Fischer, & I. Molenaar (Eds.). *Rasch models: Foundations, recent developments, and applications* (pp. 271–292). New York: Springer.

Anderson, M. (1992). *Intelligence and development: A cognitive theory*. Oxford: Blackwell Publishing.

Arden, R., & Plomin, R. (2007). Scant evidence for Spearman's law of diminishing returns in middle childhood. *Personality and Individual Differences, 42*, 743–753.

Balinsky, B. (1941). An analysis of the mental factors of various age groups from nine to sixty. *Genetic Psychology Monographs, 23*, 191–234.

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latentvariables. *Structural Equation Modeling, 12*(4), 513–535.

Brand, C. R. (1984). Intelligence and inspection time: An ontogenic relationship? In C. J. Turner, & H. B. Miles (Eds.). *The biology of human intelligence*. Nafferton: Humberside, England.

Brant, A. M., Munakata, Y., Boomsma, D. I., DeFries, J. C., Haworth, C. M., Keller, M. C., ... Wadsworth, S. J. (2013). The nature and nurture of high IQ: An extended sensitive period for intellectual development. *Psychological Science, 24*(8), 1487–1495.

Carlstedt, B. (2001). Differentiation of cognitive abilities as a function of level of general intelligence: A latent variable approach. *Multivariate Behavioral Research, 36*, 589–609.

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

Detterman, D. K., & Daniel, D. (1989). Correlations of mental tests with each other and with cognitive variables are highest for low IQ groups. *Intelligence, 13*, 349–359.

Detterman, D. K., Petersen, E., & Frey, M. C. (2001, December). Simulation of a system theory of intelligence: Critical issues. *Paper presented at the 2nd annual meeting of the International Society for Intelligence Research, Cleveland, OH*.

Detterman, D. K., Petersen, E., & Frey, M. C. (2016). Process overlap and system theory: A simulation of, comment on, and integration of Kovacs and Conway. *Psychological Inquiry, 27*(3), 200–204.

Detterman, D. K., Thompson, L. A., & Plomin, R. (1990). Differences in heritability across groups differing in ability. *Behavior Genetics, 20*(3), 369–384.

Facon, B. (2006). Does age moderate the effect of IQ on the differentiation of cognitive abilities during childhood? *Intelligence, 34*, 375–386.

Garrett, H. E. (1946). A developmental theory of intelligence. *The American Psychologist, 1*, 278–372.

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry, 27*(3), 151–177.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical manual. *Woodcock Johnson IV*. Rolling Meadows, IL: Riverside.

Molenaar, D., & Dolan, C. V. (2014). Testing systematic genotype by environment interactions using item level data. *Behavior Genetics, 44*, 212–231.

Molenaar, D., Dolan, C. V., & van der Maas, H. L. J. (2011). Modeling ability differentiation in the second order factor model. *Structural Equation Modeling, 18*, 578–594.

Molenaar, D., Dolan, C. V., & Verhelst, N. D. (2010). Testing and modeling non-normality within the one factor model. *British Journal of Mathematical and Statistical Psychology, 63*, 293–317.

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.

Murray, A. L., Dixon, H., & Johnson, W. (2013). Spearman's law of diminishing returns: A statistical artifact? *Intelligence, 41*(5), 439–451.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (Sixth edition). Los Angeles, CA: Muthén & Muthén.

Niileksela, C. R., Reynolds, M. R., & Kaufman, A. S. (2013). An alternative Cattell–Horn–Carroll (CHC) factor structure of the WAIS-IV: Age invariance of an alternative model for ages 70–90. *Psychological Assessment, 25*(2), 391–404.

Reynolds, M. R., & Keith, T. Z. (2007). Spearman's law of diminishing returns in hierarchical models of intelligence for children and adolescents. *Intelligence, 35*, 267–281.

Reynolds, M. R., Keith, T. Z., & Beretvas, N. (2010). Use of factor mixture modeling to capture Spearman's law of diminishing returns. *Intelligence, 38*, 231–241.

Rózsa, S., Kő, N., Mészáros, A., Kuncz, E., & Mlinkó, R. (2010). A WAIS–IV felnőtt intelligenciateszt magyar kézikönyve. *Hazai tapasztalatok, vizsgálati eredmények és normák*. OS Hungary: Tesztfejlesztő.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

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

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

te Nijenhuis, J., & Hartmann, P. (2006). Spearman's "law of diminishing returns" in samples of Dutch and immigrant children and adults. *Intelligence, 34*, 437–447.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology, 45*, 1097–1118.

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence, 40*, 73–76.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of cognitive abilities*. Itasca, IL: Riverside Publishing.