This article was downloaded by: [Rianne Janssen]
On: 17 October 2011, At: 09:02
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Applied Measurement in Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hame20

Measuring Mastery Across Grades: An Application to Spelling Ability

Daniël Van Nijlen [a] & Rianne Janssen [a]

[a] Faculty of Psychology and Educational Sciences, Katholieke Universiteit

Available online: 12 Oct 2011

To cite this article: Daniël Van Nijlen & Rianne Janssen (2011): Measuring Mastery Across Grades: An Application to Spelling Ability, Applied Measurement in Education, 24:4, 367-387

To link to this article: http://dx.doi.org/10.1080/08957347.2011.607064

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages

whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

APPLIED MEASUREMENT IN EDUCATION, 24: 367–387, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0895-7347 print / 1532-4818 online
DOI: 10.1080/08957347.2011.607064

Routledge
Taylor & Francis Group

Measuring Mastery Across Grades: An Application to Spelling Ability

Daniël Van Nijlen and Rianne Janssen
Faculty of Psychology and Educational Sciences, Katholieke Universiteit

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

The distinction between quantitative and qualitative differences in mastery is essential when monitoring student progress and is crucial for instructional interventions to deal with learning difficulties. Mixture item response theory (IRT) models can provide a convenient way to make the distinction between quantitative and qualitative differences in mastery. The use of latent groups, rather than focusing on manifest groupings like gender or grade, in these models is very informative to give a substantive interpretation to the qualitative differences. In the current study, mixture IRT modeling is applied to the mastery of two crucial rules in vowel duration spelling in Dutch by pupils in the four final grades of primary school. Results indicate that differences in mastery of the spelling rules are not strictly quantitative. Three latent groups of pupils can be distinguished that show qualitative differences in the mastery of one of the crucial spelling rules involved.

In student monitoring systems mastery is measured across several grades. Differences in mastery and, consequently, differences in task performance will occur and these can pose specific challenges in the data modeling. Models that deal with strictly quantitative differences (differences in degree) will not always suffice as different strategies might be applied by students when they progress through the different grades. The use of different strategies might result in qualitative differences (differences in kind) that require specific ways of dealing with.

The distinction between quantitative and qualitative differences (De Boeck, Wilson, & Acton, 2005) can be clarified as follows: quantitative differences imply

Correspondence should be addressed to Daniël Van Nijlen, Faculty of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Dekenstraat 2, P.O. Box 3773, 3000 Leuven, Belgium. E-mail: daniel.vannijlen@ped.kuleuven.be

that students differ by a degree of something identical (the latent dimension is the same; "more of the same"). Qualitative differences refer to differences that cannot be related to a degree of something identical, but something of a different kind is involved (different latent dimensions). In the context of mathematical problem solving, for example, this distinction refers to the fact that older students become better at executing certain arithmetic rules that are also used by younger students (quantitative differences), but also that they start applying different strategies, like estimation strategies (qualitative differences).

In the current paper data are analyzed for a student monitoring test for spelling and it is investigated whether progress can be modeled in a strictly quantitative way or that also qualitative differences are present. Standard models from item response theory (IRT; van der Linden & Hambleton, 1997), that reflect quantitative differences, are applied next to models that incorporate qualitative differences. Qualitative differences are linked to the topic of differential item functioning (DIF). DIF will be addressed, both using manifest groupings as latent groups. In the latter case mixture IRT models are used.

In the following, a short overview of cognitive development is given. Afterward, different ways to deal with quantitative and qualitative differences in development are presented. Finally, the spelling test used in the present study is introduced.

THEORIES OF COGNITIVE DEVELOPMENT

Strategy-use has received considerable attention in research about learning. Different types of knowledge, and hence different strategies, can be used when solving a task. How learning evolves has been addressed in a number of theories of cognitive development. In stage theories of cognitive development (Piaget, 1983; Ehri, 1992, 2000) the transition from one stage to another is characterized by a qualitative shift in the strategy that is being applied. So it is assumed that different strategies are used in different stages of development. However, according to the overlapping waves theory (Siegler, 1996) learners may use different, competing strategies simultaneously at different times of development, but along the developmental continuum there might be varying rates in the usage of each strategy. So the latter model is characterized by a gradual change in strategy use, rather than the more abrupt change in stage theories.

In stage theory as well as overlapping waves theory one expects qualitative differences to appear in task performance for children at different points of the developmental continuum (i.e., different types of strategies are being applied). Both theories, however, also can incorporate the idea of a better execution of the strategies within a stage. This increase in effectiveness will result in quantitative differences rather than qualitative differences. So, irrespective of the theoretical

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

developmental framework that is being used, the concepts of quantitative and qualitative differences are crucial.

Making a distinction between qualitative and quantitative differences in task performance is not only relevant in the theorizing about development, it is extremely relevant in the context of learning disabilities or learning difficulties (Bourassa & Treiman, 2001). It is a crucial distinction with regard to the research on the appropriate instructional interventions for certain learning disabilities or difficulties (Yang, Shaftel, Glasnapp & Poggio, 2005). If learning difficulties are merely characterized by a quantitative lag in development, a different approach will be needed then if pupils with learning difficulties show qualitative differences, i.e. they use different kinds of strategies.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

DIFFERENTIAL ITEM FUNCTIONING

From a measurement perspective, the distinction between qualitative and quantitative differences within a certain group of students can be linked to the concept of DIF. DIF refers to the presence of different measurement properties of items for different groups after controlling for the overall (latent) ability measure. If two subjects with the same ability from two groups (e.g., boys and girls; grade) have different probabilities of solving an item, this will be considered to be DIF. Purely quantitative differences in item performance between groups will not be considered to be DIF. Qualitative differences between groups in their performance on items, however, will be reflected in different measurement properties of the items and, thus, can be considered as DIF. If, for instance, students from different grades would apply different strategies in spelling certain items, these items may have a different difficulty for different grades.

In the theorizing about DIF and the praxis of DIF there has been a shift throughout the years (Zumbo, 2007). Initially, DIF was merely considered to be a nuisance to the measurement properties of the test and, hence, the focus lay on the detection of DIF items so that these items could be excluded from the test ("test cleaning"). Later on there has been a transition from pure DIF detection to also explaining why DIF was occurring and what can be learned about the item response process from the DIF present in the test. DIF turned out to be possibly interesting from a substantive point of view.

Differential item functioning may be viewed as multidimensionality in the data (Ackerman, 1992), where the different groups perform in a different way on the additional dimension(s) in the data. These secondary dimensions can be regarded as nuisance dimensions if they are not intended to be measured or as auxiliary dimensions if they are (Roussos & Stout, 1996). Hence, the presence of DIF can be checked by evaluating whether it is possible to achieve satisfactory model fit using a unidimensional IRT-model. One can also explicitly check for

the dimensionality of the data, for example by performing a factor analysis on the (polychoric) item intercorrelation matrix.

If it would be possible to fit a unidimensional IRT-model, this would mean that there is no DIF and that merely quantitative differences in mastery are present. If this is not the case, one can allow for multidimensionality in the form of DIF for some manifest, preexisting groups (e.g., gender, ethnicity, grade). In doing so the groups reflect the secondary (nuisance) dimension (Cohen & Bolt, 2005). In such a DIF-model one allows the items to have different measurement properties for different manifest groups.

In this kind of classical DIF analysis one only detects DIF for manifest groups. However, Cohen and Bolt (2005) found a weak relationship between the manifest characteristic used to investigate DIF and the actual advantaged or disadvantaged groups. Manifest groups are not as homogeneous with regard to the variable of interest as often assumed (De Ayala, Kim, Stapleton, & Dayton, 2002; Samuelsen, 2005). For instance, not all the boys will respond to the test as the "typical" boy and as a consequence the manifest grouping will not always be the most interesting grouping variable to investigate.

Tatsuoka, Linn, Tastuoka, and Yamamoto (1988) already pointed to the fact that group membership often is only a weak proxy for the aspects that are of actual instructional relevance, thus, resulting in a lack of interpretability of the results from a DIF analysis based on group membership. Differences in the instructional background can be considered to be more important determinants of potential differential performance on the items than the more evident group membership, like grade. Tatsuaoka et al. (1988) applied this idea to the analysis of a subtraction test by not using the typical DIF variables, but using the instructional method that had been used by the different teachers. This approach makes it possible to give a more substantive interpretation as to why DIF is occurring. However, the use of manifest subgroups, although based on some criterion relevant to instruction, still limits the possibility of actually detecting the most relevant distinctive criterion the can account for the DIF.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

MIXTURE IRT MODELS AND DIF

Items can function differently for certain examinees because they use different strategies to solve the item. Certain items may be very difficult when an examinee applies one solution strategy but considerably easier under another strategy (Mislevy, Levy, Kroopnick, & Rutstein, 2008). Mislevy et al. (2008) illustrate this with the distinction between the application of a rotational or an analytical strategy in a spatial visualization task. If one has to evaluate whether two representations of triangles are equal or are mirrored, one can mentally rotate the objects or analytically evaluate the shape of the triangle. Under the first strategy the amount of rotation is a crucial determinant of the difficulty of the task, while it

has no impact under the analytical strategy. The applied solution strategy is latent and often not directly linked to a manifest characteristic, so examinees need to be classified into latent groups that correspond to a certain solution strategy.

Looking for latent groups in the context of DIF has recently gained considerable attention (Cohen & Bolt, 2005; De Ayala et al., 2002; Samuelsen, 2005; Webb, Cohen, & Schwanenflugel, 2008), also in the context of modeling spelling ability (Hoijtink & Notenboom, 2004; Notenboom & Reitsma, 2007). The use of mixture IRT models (e.g., Rost, 1990; Mislevy & Verhelst, 1990) can circumvent the problems that arise with the use of manifest groups. Mixture IRT models combine a latent-class approach with an IRT-approach. The latent-class aspect of the mixture IRT model will divide the respondents into latent classes based on their response pattern, whereas the IRT aspect will make it possible to quantify performance differences within the (latent) groups of respondents. Mixture IRT models can detect subpopulations that show qualitative differences (latent classes) and at the same time quantify the differences in ability within the groups (IRT scale).

Mixture IRT models can detect latent groups for which the items function differently. First, they make it possible to investigate whether the studied population actually consists of more than one (latent) group. Second, by checking the differences in item parameters between the latent classes, one can determine whether there is some differential item functioning present in the measurement of the phenomenon of interest.

From a multidimensional perspective on DIF this can be seen as if the mixture model is used to identify the secondary (nuisance) dimension in a discretized way (Cohen & Bolt, 2005). For more information on the equivalence of multidimensional IRT models and mixture IRT models we refer to Reise and Gomel (1995) and Rijmen and De Boeck (2005).

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

AN APPLICATION: VOWEL DURATION SPELLING IN DUTCH

In this section some information is provided on the issue of vowel duration spelling for which it is investigated whether progress can be looked at in a purely quantitative way or that also qualitative differences are present.

Vowel Duration Spelling Rules

The orthographic representation of phoneme duration is a spelling problem in a number of orthographies (see Landerl & Reitsma, 2005 for an overview). A specific problem in Dutch is vowel duration spelling (i.e., the writing of long and short vowels) (Notenboom & Reitsma, 2003). The basic principle applied is that of phoneme–grapheme correspondence ("what you hear is what you write"). If you hear a long vowel you write a double letter (e.g., "l<u>aa</u>t"; *late*); if you hear a short vowel you write a single one (e.g., "l<u>a</u>t", *ruler*). However, this principle

is not sufficient for the writing of long and short vowels in words consisting of several phonological parts.

If a long vowel is followed by another phonological part it is represented with a single letter (e.g., "later"; *later*). This rule of "vowel singularization" in the remainder of the text will be referred to as the VS rule. For short vowels, the consonant following a short vowel is doubled (e.g., "ladder"; *ladder*). The use of these postvocalic consonant doublets to indicate a short preceding vowel is referred to as the "doubling consonant" (DC) rule. Often this rule is applied for the spelling of the plural of some nouns, (e.g., "lat"—"latten"; *ruler—rulers*), in derivative words (e.g., "pret"—"prettig"; *pleasure—pleasant*), or in words as such, for instance, "ladder" (*ladder*) or "terras" (*terrace*). In English a similar pattern can be found, for instance to make the distinction between *diner* and *dinner*, where the doubling of the consonant indicates that the preceding vowel is short.

Both the VS rule and the DC rule can be considered to be strictly phonological, which may cause spelling inconsistency at the morphological level as the same word stem may have a different spelling in different word forms (Landerl & Reitsma, 2005). For instance, the word stem "laat" is altered in the comparative "later." This inconsistency might pose specific problems for beginning spellers if preservation of the word stem would be a more dominant principle in early spelling development (e.g., spelling "later" as "laater"). The findings of Landerl and Reitsma (2005) support the idea that Dutch-speaking early spellers also have the tendency to apply this principle of morpheme consistency although a phonological strategy would be sufficient for an accurate spelling. For more general information on the use of consonant doublets in Dutch, also based on morphological rules (e.g., for the past-tense), we refer to the work by Notenboom and Reitsma (2007).

As is often the case in language spelling, there are some additional rules and exceptions to the general DC rule, in which case the consonant is not doubled. First, the consonant is not doubled when it is preceded by a silent vowel (referred to as SV; e.g., "monniken," *monks*). A second additional rule is that the consonant is not doubled after a diphthong (DT; e.g., "houten," *wooden*). A third additional rule states that a single consonant is written after eu, i(e) and oe (EIO; e.g., "kleuter," *toddler*). Finally, there are some words that are merely exceptions to the DC rule (e.g., "papegaai," *parrot*). Note that for these words the spelling of the stem is preserved. The stem of "houten" is "hout" (*wood*) and there is no change in the spelling of the stem for the derivative word.

Quantitative and Qualitative Differences in Spelling Ability

Strategy-use has also received considerable attention in spelling research. Both stage theories of spelling development (Ehri, 1992, 2000) and overlapping waves theory-accounts of spelling development (Rittle-Johnson & Siegler, 1999) have

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

been proposed. The question whether performance differences are merely quantitative or also qualitative differences exist can also be asked with regard to the writing of short and long vowels in Dutch. If only quantitative differences exist, it should be possible to suffice with a single unidimensional (IRT) scale. However, qualitative differences in item performance may be present. Since the rules under consideration are only taught throughout the different grades of primary school, different strategies may be used by pupils in different grades. This may cause DIF in the items across the grades. However, qualitative differences in response patterns may be attributed to known sources (manifest categories like grade) but also may be due to unknown sources (latent categories). As has been pointed out before, the manifest grouping (i.e., grades) might not always be the most relevant variable to investigate. Therefore, the presence of latent groups is checked by using mixture IRT models.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

METHOD

Test

In order to detect children that have problems with the spelling rules for the writing of short and long vowels in Dutch, a specific spelling test was constructed assessing the ability of pupils in applying these rules (Vangeneugden, 2004). The test consisted of two versions to be used in a student monitoring system. The first version was intended to be used for the detection of spelling difficulties with regard to these rules, while the second version could be administered to evaluate the progress a pupil made. An overlap in test items was included so that it would be possible to fit both tests on a common IRT-scale, in order to ensure the comparability of results and evaluate progress in a meaningful way. To achieve this combined goal of detection and progress evaluation it is necessary to also perform analyses across grades.

In the current article analyses will focus on the test version that was used for the screening of the children. The test consisted of 39 words embedded in a sentence that clarified the meaning of the word that had to be spelled by the pupils. The spelling was only scored for the application of the rules under consideration. Scoring could be done at the word level or the item level when the relevant rules had to be applied more than once within a word. The scoring at the item level resulted in a total of 50 items, based on the spelling of the 39 words.

Participants

The test was administered to 1078 children from grade 3 through grade 6 of primary education in Flanders. With regard to the language spoken at home, the

majority of these pupils was strictly Dutch-speaking ($n = 1{,}017$), 39 of them spoke Dutch in combination with another language and 22 did not speak Dutch at home. About half of the students were girls ($n = 540$).

Analyses

A scoring at the item level was used. Of the 50 items, 13 referred to the writing of long vowels (VS rule), 13 to the doubling of the consonant after a short vowel (DC rule) and 24 to the application of additional rules to the latter rule. Only pupils with a non-perfect response pattern were included ($n = 1{,}029$) in the analyses.

First, a standard IRT approach was taken in analyzing the data. Unidimensional IRT models (Rasch and the two-parameter logistic (2PL) model) were fit to the data using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). Second, the dimensionality of the data was checked by performing a factor analysis on the item intercorrelation matrix using the program TiaPlus (CITO, 2009). Next, to check for potential DIF across grades, a Rasch scale DIF analysis was done using BILOG-MG. Finally, mixture IRT models with one to four latent classes were estimated. More detail on the mixture IRT model is provided in the next section.

Formal Introduction of the Mixture IRT Model

Before presenting the mixture Rasch model, a short description of the Rasch model will be given, together with some information on the 2PL model. In the Rasch model (Rasch, 1960) the probability of solving an item is determined by the person's ability level (indicated as θ_p) and the item difficulty (indicated by β_i). In Equation 1 y_{ip} is the binary response variable for person p on item i that can either take the value 1 (correct) or 0 (incorrect).

$$P(y_{ip} = 1) = \frac{\exp\left(\theta_p - \beta_i\right)}{1 + \exp\left(\theta_p - \beta_i\right)} \tag{1}$$

If the ability level of the person equals the item difficulty the probability of solving the item will be equal to .50. If it exceeds the item difficulty this probability will be greater than .50. If the item difficulty exceeds the ability level the probability will be smaller than .50. Both the items and the respondents will be ordered on one common measurement scale, according to difficulty and ability, respectively.

In the two-parameter logistic model (Birnbaum, 1968) an additional parameter α_i for each item is included. This additional parameter reflects the different discriminative power of the items and, thus, is referred to as the discrimination parameter. The discrimination parameter represents the amount of information an

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

item gives about the latent trait, the ability of a respondent, on a certain point of the scale.

$$P(y_{ip} = 1) = \frac{\exp \alpha_i \left[(\theta_p - \beta_i) \right]}{1 + \exp \alpha_i \left[(\theta_p - \beta_i) \right]} \tag{2}$$

The mixture Rasch model (Rost, 1990) is an extension of the Rasch model that adds a class weight π_g to the equation. This class weight reflects the probability of belonging to class g and is also known as the mixing parameter. The ability parameter and the item difficulty (i.e., easiness) parameter get an extra subscript to indicate that they are allowed to differ across the latent classes: θ_{pg} is the ability of person p in class g; β_{ig} is the item easiness parameter for item i in class g. So for each latent class a different measurement scale is constructed.

$$P(y_{ip} = 1) = \sum_{g=1}^{G} \pi_g \frac{\exp \left(\theta_{pg} + \beta_{ig} \right)}{1 + \exp \left(\theta_{pg} + \beta_{ig} \right)} \tag{3}$$

Furthermore, we assume $\theta_{pg} \sim N\left(0, \sigma^2\right)$ for all g. Since the class weights have to sum to one, π_G is set to$1 - \sum_{g=1}^{G-1} \pi_g$.

The models were estimated using an adaptation of the EM-algorithm with a logistic link function, introduced in Rijmen and De Boeck (2003). Estimation was done using MATLAB. Mixture IRT models can also be estimated in more specific software like WINMIRA (von Davier, 2001) or in LCA-oriented software like Latent GOLD (Vermunt & Magdison, 2000).

RESULTS

First the results of the psychometric analyses will be presented. Second, the interpretation of the results for the selected model will be addressed.

Psychometric Analyses

Several models were fitted to the spelling data. The different models will be compared with regard to fit and interpretability and the selection of the most appropriate model will be discussed.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

Non-mixture IRT models. No model-data fit could be obtained for the total group, neither for the Rasch model nor for the less restrictive 2PL. The total chi-square value in BILOG for the Rasch model was 1987.2 with 350 degrees of freedom ($p < .0001$). For the 2PL model the chi-square value was reduced considerably ($\chi^2 = 452.4$; $df = 351$; $p = .0002$), but still the model did not fit. Although the 2PL's overall fit was much better, fit at the item level still was not satisfactory. Moreover, some parameter estimates were not really interpretable from a substantive point of view. For almost all the items pertaining to the additional rules to the DC rule, the item difficulty parameters were extremely low in combination with a very low discrimination parameter. This can be indicative of multidimensionality in the data and suggest that these items are tapping a different underlying dimension. As also these items are essential for the content validity of the test, the 2PL did not lead to a useful model for the data.

Analyses were not only performed across grades, but also per grade. For all grades the Rasch model did not fit, but the 2PL did fit according to the chi-square test. However, given the limited number of students per grade fit can be easily obtained without actually getting a meaningful result. Moreover, again the results for the 2PL model were very difficult to interpret in a meaningful way because of extremely low difficulty parameters in combination with very low discrimination parameters for the items pertaining to the additional rules.

Dimensionality of the data. The dimensionality of the data was investigated by performing an item factor analysis, both across grades and per grade. A polychoric item intercorrelation matrix was calculated for each group and a factor analysis was done. The scree-plot indicated that for the total group at least two dimensions could be distinguished. The analyses per grade indicated three dimensions for grade 3 and 6 and two dimensions for grade 4 and 5.

The item factor analyses clearly showed that the spelling data were not unidimensional. Possibly this could be due to DIF for students from different grades, which was investigated in the next step, but the analyses per grade already hinted at the fact that some other process was involved.

DIF-analysis for grade. As mentioned before, the fact that the standard IRT models did not fit might be caused by the fact that children acquire these rules throughout the different grades of primary school and, as a consequence, that children in different grades apply different strategies to solve the items. A DIF-analysis using BILOG-MG indeed showed that a considerable number of items suffered from DIF across grades. DIF for an item was reflected in a differing item parameter for the four grades, after the differences in overall ability level between

grades were taken into account. However, it was fairly difficult to detect a clear, interpretable pattern in the DIF items.

In the present case, in Flanders elementary schools start teaching the rules at different times in primary school, and, children might learn about the rules at different points during primary school. Moreover, not all children will acquire the rules at the same pace and will apply them at different rates. This makes it plausible that grade might not be the most relevant variable to study and that it could be more interesting not to work with pre-defined groups. As stated before, mixture IRT models might provide a solution to this issue.

Mixture IRT models. Table 1 describes the results for the estimated mixture IRT models. For each model the loglikelihood and the number of estimated parameters is reported. The total number of parameters equals the number of items for each class, the variance of the ability parameter theta (which is constant over classes), and the mixing parameters for models with more than one latent class. Note that only G-1 mixing parameters are estimated because they are set to sum to one.

Both the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) were used for model selection. Both of these criteria penalize model complexity by taking into account the number of estimated parameters and provide relative information on the model fit. The model with the lowest value for the information criterion is preferred. Also, the size for each of the latent classes is reported.

Both the AIC and BIC indicate that a solution that assumes that all pupils belong to the same latent class is not desirable to describe the data. This was already clear from the fact that it was not possible to fit a unidimensional IRT scale to the data. Also the solution assuming that four latent classes could be distinguished was not suited for the data. The additional latent class only contained one percent of the pupils and both the AIC and the BIC indicated that this was not the preferable solution. The lowest value for the AIC indicates that a solution with

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

TABLE 1
Results Mixture Models With One to Four Latent Classes

Number of Latent Classes	*Loglikelihood*	*Number of Parameters*	*AIC*	*BIC*	*Proportion in Latent Class*			
					1	*2*	*3*	*4*
1	−16741	51	33584	33836	1.00	—	—	—
2	−15709	102	31622	32126	.36	.64	—	—
3	−15547	153	31400	32155	.23	.37	.40	—
4	−15498	204	31404	32411	.23	.01	.36	.40

three latent classes describes the data the most adequately, while the BIC refers to a solution with two latent classes as the preferable one.

For completeness the AIC and BIC were also calculated for the Rasch model and 2PL model from the analyses using BILOG-MG. For the Rasch model these values were very close to those obtained in the mixture analysis, namely 33591 for the AIC and 33843 for the BIC. For the 2PL model these values were 31786 and 32285, respectively. This means that the 2PL was still outperformed by the mixture models with at least two latent classes.

To decide on the issue of working with a solution with two or three latent classes the shift from the pupils between the two solutions was investigated. For each pupil a probability was estimated for belonging to one of the latent classes. Each pupil was allocated to the latent class with the highest probability. To facilitate interpretation the latent classes were ordered based on the average total score of pupils within a class. Compared to the two-class solution an additional middle group is created in the three-class solution. As the latent classes are ordered, one can really consider this additional group as a middle group. None of the pupils in the low-performing class of the two-class solution makes a shift to the highest-performing group in the three-class solution, but about one third of them makes the shift to this middle group. There is a similar, but opposite pattern for the pupils in the high-performing group of the two-class solution.

Given this information and given the fact that the difference between a two- and a three-latent class solution was not that big, a three-class solution was obtained, which was also selected by the AIC. Despite the fact that AIC is sometimes reported to have a tendency to overestimate the number of latent classes (Nylund, Asparouhov & Muthén, 2007), it was chosen to let interpretability be the most important factor in deciding on the number of classes.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

Interpretation

To give a substantive interpretation to the results, two approaches were used. First, the results for the latent classes were linked to the evident, manifest categories of grade and gender. This mapping of manifest groups onto the latent classes (Samuelsen, 2005) may hint at a first possible interpretation of why it was not possible to analyze the data in a strictly quantitative way and what the source of DIF is. Second, the results were interpreted based on a comparison of the item parameters for the three latent classes.

Interpretation based on classification of students. With regard to the distribution for gender across the different latent classes, there were only minor differences between boys and girls, with slightly more girls in the highest performing group and slightly less in the lowest performing. Class membership did not seem to be very much related to gender.

TABLE 2
Cross-Tabulation of Grade and Class Membership: Proportion Calculated Row-Wise

	Class			
Grade	*1*	*2*	*3*	*Total*
3	.60	.32	.08	267
4	.18	.46	.36	262
5	.09	.39	.52	270
6	.02	.30	.68	230
Total	235	381	413	1029

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

The latent classes clearly differed in their composition when grade is considered (Table 2). The majority of grade 3 (60%) belonged to the first, lowest performing latent class, while the majority of grade 6 (68%) belonged to the third, best performing class. This result suggests that indeed the lack of fit for a unidimensional model might be due to different strategy usage of pupils dependent on the extent to which the rules are mastered during their years in primary school.

However, grade at best can be considered as a partial proxy for class membership, but not a very accurate one. For instance, 8% of the grade 3 pupils in fact belonged to the third, best-performing, latent class and for each grade at least about one third of the pupils is classified within the second, medium-performing class. This result, in its turn, suggests that acquiring these rules is a more subtle process than one that can be captured by grade. An interpretation based on the item parameters for the latent classes might give us more insight in what actually constitutes the typicality of each latent class.

Interpretation based on item parameters: Significance testing. The item parameters for the three latent classes were compared using one reference class (i.e., the lowest performing class) and correcting the item parameters of the other classes for the average performance difference with this reference class (DuToit, 2003). As the average ability for each class was set to 0, the overall performance difference for the three classes was reflected in the difference of the average item easiness (i.e., the higher the average item easiness, the higher the average overall ability in vowel duration spelling). Then the difference in average item easiness with the reference class was subtracted from the item parameters for the other two classes to get "equated" item parameters.

The significance of differences in item difficulties for the latent classes was tested, taking into account the standard errors for these parameters. If the corrected item parameters are not significantly different from each other, this indicates that there is no DIF. The presence of significant DIF was tested for each pair of latent classes.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

Several situations can occur for each item depending on which of the pairwise comparisons of latent classes are significant. The simplest case is if none of the comparisons are significant, pointing to the fact that there is no DIF for the item. Other clear situations are when the items, going from the low performing (LP) over the medium performing (MP) to the high performing (HP) group, become gradually easier (HP > MP > LP) or more difficult (LP > MP > HP). Sometimes one of the steps from one group to the other is not significant but they still follow a gradual pattern (e.g., LP = MP > HP). For a few items the results for the significance tests were not in accordance with some logical pattern. For instance, when the item is easier for the medium performing, but there is no difference between the low performing and the high performing group. These are combined under the label "other."

The results are presented in Table 3. The items concerning the writing of long vowels (VS rule) actually showed almost no DIF across the latent classes. Almost all of the items measuring the DC rule turned out to become gradually easier for the latent classes. The results for the additional rules to the DC rule seemed to be less straightforward to interpret. Six of the items pertaining to the additional rules showed no DIF at all between the three groups. However, five out of these six items referred to one of the specific additional rules (EIO). If we disregarded these items, almost all of the remaining additional rule items became gradually more difficult for the latent classes.

Interpretation based on item parameters: Visual inspection. A visual representation of the corrected item parameters, besides the significance testing, can provide crucial information and facilitate a substantive interpretation. In Figure 1 this visual representation of the corrected item parameters is presented

TABLE 3
Results Significance Tests Pairwise Comparison of Item Parameters for the Latent Classes

	Vowel Singularization Items	*Doubling Consonant Items*	*Additional Rules and Exceptions*
HP > MP > LP	0	10	0
HP = MP > LP	1	1	0
HP > MP = LP	0	1	0
LP > MP > HP	0	0	8
LP > MP = HP	0	0	3
LP = MP > HP	1	0	6
Other	3	1	1
No DIF	8	0	6
Total	13	13	24

HP = high-performing class; MP = medium-performing class; LP = low-performing class.

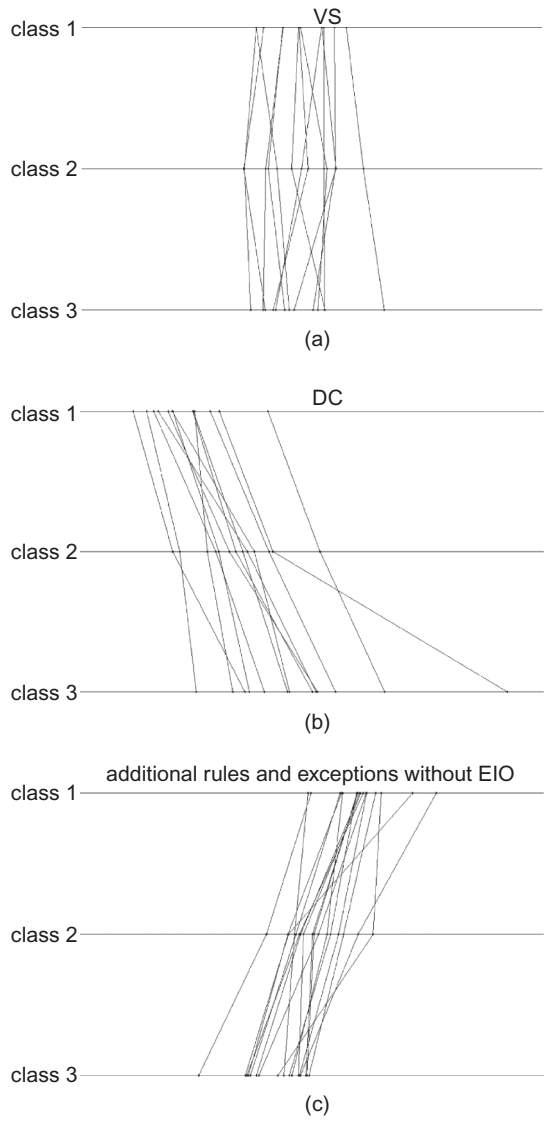

FIGURE 1 Graphical illustration of the corrected item easiness parameters for the latent classes. Top panel (a) VS rule items, middle panel (b) DC items, and bottom panel (c) items for the additional rules to the DC rule (except for items pertaining to the EIO rule). Each line represents the measurement scale for a latent class. Top line represents item easiness parameters for the low performing latent class, middle line for the medium performing latent class, and bottom line for high performing latent class.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

in three panes. A separate plot was made for the VS rule items, the DC items and the items pertaining to the additional rules to the DC rule, disregarding the EIO-items. As for each item an item easiness parameter was estimated the difficult items are represented at the left-hand side of the scale and the easier items on the right-hand side.

The top line of the plot represents the item parameters for the low performing class. On the middle line the item parameters for the medium performing latent class are plotted. On the bottom line the item parameters for the high performing latent class are represented. Vertical lines connect the parameters for the same items in the different latent classes. If there would be no DIF in any of the items (and there would be no error in the parameter estimates) these would be straight, vertical lines.

For the items pertaining to the VS rule (Figure 1a) the lines do go down in a rather straight vertical way, indicating that there is not really an issue with DIF for these items. Moreover, these items seem to be fairly easy for all three latent classes. In Figure 1b the DC items are plotted and there is a clear tendency in the vertical lines. The DC items are difficult for pupils in the first latent class, become a bit easier for the ones in the medium performing latent class and are clearly easier for the pupils in the high-performing latent class. For the items pertaining to the additional rules (Figure 1c) we see an opposite pattern in the vertical lines. There is a tendency for these items to be relatively easy for the pupils in the low performing latent class and they become gradually more difficult for the higher performing latent classes.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

DISCUSSION

Data with regard to two crucial spelling rules in Dutch were analyzed, applying mixture IRT models. From these analyses it was clear that pupils from grade 3 to grade 6 could not be considered to be one homogeneous group with regard to the mastery of these rules. Not only were there quantitative differences in their performance on the test, but it was also necessary to take into account qualitative differences.

Interpretation of the Latent Classes

Three latent classes could be distinguished, primarily based on the mastery of the rule for doubling consonants following short vowels. The first class consisted of low-performing spellers that did not (yet) master the “double consonant” rule but, as a consequence, have relatively little problems with the items concerning the additional rules and exceptions when the DC rule should not be applied. The second class contained medium-performing pupils who are still acquiring the DC

rule. And finally, the third group was made up of high-performing spellers for whom the DC rule is relatively easy but who overgeneralize this rule and apply it in cases they should not use it.

The relative easiness of the additional rules and the exceptions for the first class can be explained in two ways. First, in general, a simple strategy of phoneme–grapheme translation, a sound-based approach, for these items will result in a correct spelling (Landerl & Reitsma, 2005). This result is in line with Notenboom and Reitsma (2007) who found supporting evidence for the hypothesis that on the onset of development children predominantly apply an alphabetic, sound-based strategy. A second explanation only pertains to plurals and derivative words. For these words, when the additional rules apply, the spelling of the word stem is preserved so the principle of morpheme consistency is not violated (e.g., "monniken" with the stem "monnik"). The findings of Landerl and Reitsma (2005) suggest that for early spellers this is a dominant principle. These two explanations also account for the relative difficulty of the DC items, for which a sound-based approach results in an incorrect spelling and, for the plurals and derivative words, the spelling of the word stem is not preserved (e.g., "prettige" for which the word stem is "pret").

The second latent class can be considered to be some kind of transition group of pupils that start using consonant doublets at times, resulting in a relatively better performance on the DC items, but a decrease in the correct answers for the additional rules and the exceptions. From the perspective of the overlapping waves model, this might be considered to be a group of pupils that combine the sound-based approach of the first group with a rule-based approach that might be more typical for the third latent class.

The relative difficulty of the items pertaining to the additional rules and the exceptions for the pupils in the third class points to overgeneralization of the use of the postvocalic consonant doublet in vowel duration spelling. Overgeneralization or "overuse" of some strategies or rules (Bourassa & Treiman, 2001) is a commonly observed issue in language acquisition. One example is the overuse of the suffix -ed for the past tense in English (e.g., spelling slept as "sleped"; Nunes, Bryant, & Bindman, 1997). Also Notenboom and Reitsma (2007) point to the occurrence of overgeneralization resulting in a U-shaped developmental pattern in the context of spelling doublets in Dutch. This means high scores at the onset, a decrease in the scores due to overgeneralization of related rules and then an increase in the scores when the overgeneralization no longer occurs. They list a few studies that found this kind of developmental pattern in other alphabetic languages like Danish, Greek, French, English, and Brazilian Portuguese.

For the VS rule the pupils gradually become better at applying the rule, in line with their general progress, as was reflected in the limited number of DIF occurrences for these items. For these items almost only quantitative differences

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

between the classes seem to occur, pupils from the different latent classes just show a different effectiveness in the execution of the rule. A comparable interpretation can be given to the EIO items that behave somewhat differently from the other exceptions to the DC rule.

Relevance of the Applied Methodology

The application of mixture IRT models to these educational data made it possible to reveal a pattern that could not be detected using a classical DIF-procedure. Qualitative differences could be detected between non-manifest groups of pupils and simultaneously quantitative differences between pupils could be captured. This example illustrates that the application of this kind of models might have great value in an educational context to find out where specific difficulties in acquiring a domain can be found.

The results of the present article, for instance, make it possible for specialists in spelling didactics to look into the process of overgeneralization in the acquisition of spelling rules and think about possible ways to deal with it. At the same time, from a diagnostic point of view, the use of these models enables one to not only detect quantitative differences in mastery of specific educational issues, but also to find out in a qualitative way where the specific problems of a pupil lie. Based on their response pattern, pupils can be classified with a certain probability in one of the latent groups that show specific difficulties in vowel duration spelling.

Other areas of application of mixture IRT models include personality assessment (Reise & Gomel, 1995) where qualitative variations in the personality trait structure can be modeled using mixture models (e.g., distinct types of depression). Also for the modeling of speededness effects in time-limit tests mixture IRT models can be applied (Bolt, Cohen & Wollack, 2002) or, somewhat related, for the modeling of (random) guessing behavior with multiple choice items (Mislevy & Verhelst, 1990; von Davier & Yamamoto, 2007). In both cases the different latent classes refer to groups that apply different strategies to solve items (speeded versus non-speeded respondents and guessers versus non-guessers).

Limitations of the Present Study

Obviously there are some limitations to the present study. However tempting a developmental interpretation might be, we should be aware that, strictly speaking, these (cross-sectional) data do not provide the necessary information for this. It is not necessarily so that an individual pupil progresses from the first latent class to the second and further on to the third. The differential composition of the latent classes for grade might hint at some developmental pattern in the latent classes, but it is no conclusive evidence. To make this type of interpretation possible longitudinal data should be collected.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

Still, it is worth noting that no extra latent class could be distinguished of pupils that master the DC rule, but at the same time do not overgeneralize the rule to instances where the rule does not apply. This would be reflected in a class with comparable item easiness parameters to the third class for the DC items and comparable item parameters to the first latent class for the additional rules and the exceptions. It might be hypothesized that if older children were included in the sample this kind of latent class would have appeared. It makes sense that the spelling development does not finish by the end of primary education and the extension of the age group may result in a more complete picture.

Another qualification to the results is that the interpretations about the approach used to solve an item are post-hoc, they are not based on observations, self-reports of the pupils or the like. Moreover, the focus of this study did not lie on specific minority groups or children with specific learning difficulties. As a consequence limited background information was gathered in the data collection. The inclusion of more background information in the analyses of the data could be a fruitful elaboration of the current research.

An obvious limitation of the study pertains to the fact that it only concerns data on a specific spelling issue (i.e., vowel duration spelling) in one language (i.e., Dutch). However, it was illustrated that a more general developmental pattern was elucidated, namely that of overgeneralization in language acquisition. Moreover, distinguishing quantitative differences from qualitative differences is relevant in a wide variety of educational contexts. Hence, the relevance of the applied methodology is not limited to this one specific spelling issue in one not so widespread language.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

ACKNOWLEDGMENTS

The present research was supported by the OBPWO grant 01.08 and the 2006–2011 national assessments research grant of the Department of Education of the Ministry of the Flemish Community (Belgium). We thank Evi Vangeneugden for collecting the data used in this study and the initial analyses on the data. Furthermore, we thank Frank Rijmen for his assistance in analyzing the data.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Akaike, M. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under the conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.

Bourassa, D. C., & Treiman, R. (2001). Spelling development and disability: The importance of linguistic factors. *Language, Speech and Hearing Services in Schools*, *32*, 172–181.

CITO (2009). *TiaPlus* (computer program and manual). Arnhem, The Netherlands: CITO.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133–148.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 3– 4*, 243–276.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, *112*, 129–158.

DuToit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Ehri, L. C. (1992). Review and commentary: Stages of spelling development. In S. Templeton & D. R. Bear (eds.), *Development of orthographic knowledge and the foundations of literacy: A memorial festschrift for Edmund H. Henderson* (pp. 307–332). Hillsdale, NJ: Erlbaum.

Ehri, L. C. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in Language Disorders*, *20*, 19–36.

Hoijtink, H., & Notenboom, A. (2004). Model based clustering of large data sets: Tracing the development of spelling ability. *Psychometrika*, *69*, 481–498.

Landerl, K., & Reitsma, P. (2005). Phonological and morphological consistency in the acquisition of vowel duration spelling in Dutch and German. *Journal of Experimental Child Psychology*, *92*, 322–344.

Mislevy, R. J., Levy, R., Kroopnick, M., & Rutstein, D. (2008). Evidentiary foundations of mixture item response theory models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 149–175). Greenwich, CT: Information Age Publishing Inc.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195–215.

Notenboom, A., & Reitsma, P. (2003). Investigating the dimensions of spelling ability. *Educational and Psychological Measurement*, *63*, 1039–1059.

Notenboom, A., & Reitsma, P. (2007). Spelling Dutch doublets: Children's learning of a phonological and morphological spelling rule. *Scientific Studies of Reading*, *11*, 133–150.

Nunes, T., Bryant, P., & Bindman, M. (1997). Morphological spelling strategies: Developmental stages and processes. *Developmental Psychology*, *33*, 637–649.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modelling: A Monte Carlo simulation study. *Stuctural Equation Modeling: A Multidisciplinary Journal*, *14*, 535–569.

Piaget, J. (1983). Piaget's theory. In P. Mussen (Ed.), *Handbook of child psychology* (pp. 103–128). New York, NY: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reise, S. P., & Gomel, J. N. (1995). Modeling qualitative variation within latent trait dimensions: Application of mixed-measurement to personality assessment. *Multivariate Behavioral Research*, *30*, 341–358.

Rijmen, F., & De Boeck, P. (2003). A latent class model for individual differences in the interpretation of conditionals. *Psychological Research*, *67*, 219–231.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011

Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture-Rasch model. *Psychometrika*, *70*, 481–496.

Rittle-Johnson, B., & Siegler, R. S. (1999). Learning to spell: Variability, choice and change in children's strategy use. *Child Development*, *70*, 332–348.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis Paradigm. *Applied Psychological Measurement*, *20*, 355–371.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation, University of Maryland.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York, NY: Oxford University Press.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement*, *25*, 301–319.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Vangeneugden, E. (2004). *De verdere ontwikkeling van de criteriumtoets SPOMEK voor het lager onderwijs* [The further development of the criterion test SPOMEK for primary education]. Unpublished master's thesis. Department of Educational Sciences, Katholieke Universiteit Leuven, Belgium.

Vermunt, J. K., & Magdison, J. (2000). *Latent GOLD*. Belmont, MS: Statistical Innovations.

von Davier, M. (2001). WINMIRA 2001. *Latent Class Analysis, Dichotomous and Polytomous Rasch Models*. St. Paul, MN: Assessment Systems.

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–115). New York: Springer.

Webb, M. L., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test-III. *Educational and Psychological Measurement*, *68*, 335–351.

Yang, X., Shaftel, J., Glasnapp, D., & Poggio, J. (2005). Qualitative or quantitative differences? Latent class analysis of mathematical ability for special education students. *The Journal of Special Education*, *38*, 194–207.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [computer program]. Chicago, IL: Scientific Software.

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.

Downloaded by [Rianne Janssen] at 09:02 17 October 2011