Tracking educational progress
Brinkhuis, M.J.S.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 3

Detecting differential development using filters

Summary

The amount of data available in the context of educational measurement vastly increased in recent years. Such data is often incomplete, involves tests administered at different time points and during the course of many years, and can therefore be quite challenging to model. In addition, intermediate results like grades or report cards being available to pupils, teachers, parents and policy makers likely have an influence on performance, which adds to the modeling difficulties. We propose the use of simple data filters to obtain a reduced set of relevant data, which allows for simple checks on the relative development of persons, items, or both.
Chapter 3. Detecting differential development

Table 3.1: Complete data

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: Missing data

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>0</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.</td>
<td>.</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1</th>
<th>0</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

3.1 Introduction

3.1.1 Data representation

The amount and structure of data used in educational measurement has changed in the last decades. Most of test theory, such as Rasch (1960), Choppin (1968) or Lord and Novick (1968), has been developed assuming a complete data matrix, i.e., without missing data, as represented in Table 3.1. The applications of item response theory (IRT) extended such that not all persons answer all items, and not all items are answered by all persons. One example is a simple anchor test design as shown in Table 3.2. Here two test booklets with an overlapping subset of items are presented to two groups of candidates, therefore the data matrix contains structural missing values. Many ways are devised to deal with missing data in this context, such as concurrent calibration, different ways of equating, etc. (Dempster, Laird, & Rubin, 1977; Hanson & Béguin, 2002; Verhelst, 2004). While Table 3.2 provides a simple example, numerous applications exist where designs are more complex, involving many test booklets administered to many candidates. Complexity is however also added on a conceptual level, where different booklets are administered at different moments in time, e.g., over different years, to different populations, across changing implementations of educational policies and any combination of the aforementioned. Examples of complex designs are computer adaptive testings (CATs) and on-line learning environments such as computer adaptive practice (CAP) or computer adaptive learning (CAL) environments (e.g., Wauters, Desmet, & Van den Noortgate, 2010; Klinkenberg, Straatemeier, & van der Maas, 2011), which can generate sparse data like those shown in Table 3.3. A common and efficient format of presenting sparse data matrices is a so-called long format, as presented in Table 3.4, where we have an index for persons, \( p \) in our example, an index \( i \) for items, and \( x \) for the scored responses. Since many applications are sequential or administered in real time, an ordering or time index \( t \) is added. The methodology applied to these data sets involving all sorts of change, often
It still involves assumptions that are more suited to the situation presented in Table 3.1 and 3.2, such as invariant model parameters. Especially if tests are designed to inform educators how pupils, schools, or the whole educational system is developing, it is likely that this information will evoke change at certain levels of the educational system. For example, if a teacher is provided with information that his or her class is performing poorly on algebra, he or she might adapt the style and content of teaching. Hopefully, this results in a better class performance and if so, such changes might invoke a change in model parameters, for example inducing differential item functioning (DIF) (e.g., Bechger & Maris, 2014) over time, often referred to as item drift. Especially when item banks are considered (Wright & Bell, 1984; Hambleton, 1986), item difficulties are estimated and fixed. While several tests for DIF over time exist, their use for monitoring item difficulties as data cumulates is often limited.

Feedback might be regarded as a hidden Markov model (Rabiner & Juang, 1986; Visser, 2011) applied to Table 3.4:

\[
(\theta_t, \delta_t) \downarrow \hat{\theta}_t, \hat{\delta}_t \rightarrow \theta_{t+1}, \delta_{t+1}
\]  

(3.1)

where the estimated model parameters at time point \( t \) have an influence on the true parameters at time point \( t + 1 \). A hidden Markov model or any other model that attempts to model these changes over time will obviously be complex, given the vast amount of different actors, such as teachers, schools, but also variables such as gender, school year, etc., which all need to be modeled leading to intricate non-linear dynamics.

It is not attempted to define such a model here, but it serves to illustrate that educational measurements can be invasive, especially if their goal is actually inducing change. The result is a complex dynamic system with many actors.
We will introduce an algorithm that takes a different approach, and allows for testing of differential development of persons and items using filtered data.

### 3.1.2 Filtered data

Many applications in educational measurement produce the type of data presented in Table 3.4, e.g., learning environments, CAT, etc. Several learning environments and their modeling challenges are discussed in (Wauters et al., 2010). A specific example discussed in section 3.3.1 is the Math Garden, an on-line CAP environment for mathematics ability (Klinkenberg et al., 2011), includes over 220 million responses, 500 thousand daily, produced by over 30 thousand students from 900 schools. While these examples might seem quite large compared to the current day standard, they are quite small if we compare them to the potential of national student monitoring systems, such as the Cito monitoring system, which serves in the Netherlands alone close to 1.5 million students.

Due to the size and type of these (potential) applications, modeling is increasingly difficult. However, with much data at hand it can also be less necessary. With quite simple mechanisms we can filter out some of the relevant information, the strength of such an approach is in the amount of data it can handle efficiently compared to modeling strategies. It will be shown that algorithms on this filtered data can pick up differential development over time of both persons and items. In addition, the detection of differential development based on gender, birth month, item type, etc., is a simple extension of the algorithm.

### 3.2 Methods

We introduce a method to test for differential development of pairs of items, persons, or both. The basic units we consider are consecutive pairs of items or persons. In longitudinal data as presented in Table 3.4, these represent two consecutive rows where either the person index $p$ is stable, i.e., an item pair, or the item index $i$ is stable, i.e., a person pair.

Quite some works in educational measurement discusses how psychometric data can be regarded as pair-wise comparisons, where a response is generated by a person-item interaction (e.g., Guttman, 1946; Choppin, 1983; van der Linden & Eggen, 1988; Zwinderman, 1995). The idea of data filtering as mentioned in section 3.1.2 is not new in paired comparisons, specifically in the prevalence of ties in item pairs. Concerning person pairs, Choppin (1983)
already mentioned that it is theoretically possible to compare person pairs, just like comparing item pairs, but found little practical application in his days. In addition, the works that suggest adaptively calibrating item parameters using longitudinal data (Choppin, 1968; van der Linden & Eggen, 1986; Kingsbury, 2009), hold onto the assumption of invariant item parameters over time. The methods presented hereafter are however both useful for testing differential or idiosyncratic development of item pairs and person pairs.

For either an item pair or a person pair, four possible outcomes are possible in a simple dichotomous case. In Table 3.5 these outcomes are presented for an item pair. The possible outcomes include making only one of the two items correct, making both incorrect, and making both correct. A variable $Z_{ij}$ is introduced to classify the outcomes of these item pairs. Ties are indicated with $Z_{ij} = 1/2$, 1 indicates the correctness of $i$ over $j$ and 0 the correctness of $j$ over $i$ (van der Linden & Eggen, 1986). $Z^*$ is introduced as a simple reparametrization of $Z$ that allows for an easy use of cumulative sums, as will be discussed later. Though an item pair is considered in Table 3.5, the same is applicable to a person pair, only changing the column names to $X_{pi}$, $X_{qi}$, $Z_{pq}$ and $Z^*_{pq}$.

We add two assumptions to the data vector $Z_{ij}$. First, we assume a Rasch model (RM) (Rasch, 1960) model holds. Second, we assume for item pairs that the ability of person $p$, and for person pairs the difficulty of item $i$, do not change between two consecutive item responses, while change is allowed between these pairs. Given these assumptions, only non-ties are informative concerning the difference in either item difficulties or person abilities (Choppin, 1968), and therefore ties can be regarded ignorable.

For item pairs, the probabilities of answering item $i$ correct and item $j$ incorrect according to the RM are:

\[
P(X_{pi} = 1) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)} \tag{3.2}
\]

\[
P(X_{pj} = 0) = \frac{1}{1 + \exp(\theta_p - \delta_j)} \tag{3.3}
\]

\[
\begin{array}{cccc}
X_{pi} & X_{pj} & Z_{ij} & Z^*_{ij} \\
0 & 0 & 1/2 & 0 \\
0 & 1 & 1 & 1 \\
1 & 0 & 0 & -1 \\
1 & 1 & 1/2 & 0 \\
\end{array}
\]

Table 3.5: Possible outcomes for a pair of items $i$ and $j$. 
Since the formula for the probabilities for person pairs is similar, it is left out. Since we disregard ties, the probability of this specific response pattern given the two possible non-tie outcomes is for item pairs:

\[
P(X_{pi} = 1, X_{pj} = 0 | X_{pi} \neq X_{pj}) = P(Z_{ij} = 0) = \frac{\exp(\theta_p - \delta_i)}{\exp(\theta_p - \delta_i) + \exp(\theta_p - \delta_j)}
\]

if we assume that \( \theta_p \) does not change between answering \( X_{pi} \) and \( X_{pj} \). Hence, we can leave out the time index for clarity. Clearly, this probability does not depend on \( \theta_p \) anymore. It follows that under the RM, \( Z_{ij} \) is a Bernoulli process, i.e., an independent and identically distributed (i.i.d.) Bernoulli distributed sequence:

\[
Z_{ij} \sim \text{Bernoulli}(\delta_{ij})
\]

with length \( n \), i.e., the number of non-ties in \( Z_{ij} \), and \( \delta_{ij} \) being the success parameter, which can also be stated as:

\[
Z_{ij} | Z_{+,n} \sim \text{Uniform}(0, 1) \quad \forall \; i, j, i \neq j,
\]

i.e., every sequence of \( Z_{ij} \) is equally likely given lengths \( n \) and numbers of successes \( Z_{+,n} \).

Many different methods are available to test these assumptions on all item pair sequences \( Z_{ij} \), of which two related methods are discussed in this paper. First, a graphical presentation of \( Z_{ij}^* \) as a Brownian bridge is discussed, where as a reference Kolmogorov bounds are added. Second and related, the Kolmogorov-Smirnov (K-S) test statistic is used.

Possible applications are numerous, since the method can be used for differential development in person pairs, such as detecting differences in ability from reference groups, cheating, subpopulation drift, etc. For item pairs applications include DIF detection, monitoring drift in item banks, detecting non-uniform development of item difficulties due to curriculum changes, etc. Since the person pair sequences are independent of the item pair chains, both can be used jointly.
3.2 Methods

$$X_{pi} \quad X_{qj} \quad X_{pk} \quad X_{qk} \quad Z_{ij} \quad Z_{ij}^*$$

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

Table 3.6: Possible outcomes of non-ties for items $i$, $j$ and $k$.

### 3.2.1 Extended filtering

The aforementioned filter in Table 3.5 serves as a simple example of many different kinds of possible filters. Though it can be applied to large data applications in a straightforward manner, the amount of pair-wise observations of pairs of items or persons can be quite small still. Especially in applications involving some form of adaptivity, the item selection algorithm obviously increases the amount of observations for some pairs, at the cost of fewer observations of other pairs. For example, the number of pair-wise observations between a rather difficult and rather easy item involving the same person will be very limited in data from CATs. Algorithms to control item exposure provide another example where certain item-pairs are unlikely to be observed.

We provide a straightforward extension of the filter in Table 3.5 to increase the number of observations available after filtering. The size of this increase will strongly depend on how the data is generated, and an example on the efficiency of these filters is provided in section 3.3.

Consider the outcomes of two persons $p$ and $q$ responding to three items $i$, $j$ and $k$, where person $p$ only responds to items $i$ and $k$, and person $q$ to items $j$ and $k$. The outcomes that are non-ties are shown in Table 3.6. The conditional probability of observing one of these patterns is denoted as follows:

$$P(Z_{ij} = 0) = \frac{\exp(\theta_p - \delta_i + \theta_q - \delta_k)}{\exp(\theta_p - \delta_i + \theta_q - \delta_k) + \exp(\theta_q - \delta_j + \theta_p - \delta_k)}$$

$$= \frac{\exp(-\delta_i)}{\exp(-\delta_i) + \exp(-\delta_j)}$$

(3.7)

which is clearly identical to (3.4). Therefore, this extension allows to indirectly compare two item responses by two different persons, by introducing a common item $k$. The extended filter can be used in addition to the direct comparisons by the regular filter, and can contribute to a considerable amount of comparisons, as will be demonstrated in section 3.3. The extended filter in Table 3.6 demonstrates how indirect comparisons can be used too.

The assumption of this specific extended filter is that during the administration of items $i$, $j$ and $k$ to persons $p$ and $q$ a RM holds with abilities $\theta_p$ and
and difficulties $\delta_i, \delta_j$ and $\delta_k$ being invariant. Clearly, this extension can, in addition to its use in item pairs as in Table 3.6 and (3.7), also be used to enrich the amount of observation on person pairs. In section 3.3.1, this extended filter is used in a data example.

### 3.2.2 Random walks

Under the RM, $Z_{ij}$ are i.i.d. Bernoulli processes for each pair of $i$ and $j$, and thus the recoded paired responses $Z^*_{ij}$ from Table 3.5 are Markov chains, specifically random walks, for sequential observations of each pair of items. The drift parameters clearly depend on the differences in item difficulties between each pair of items. The same holds obviously when considering person pairs.

It is of interest to evaluate for each of the item pairs whether they are indeed random walks. Regardless of the method of evaluation, it allows to test for specific kinds of misfit of the RM. For example, if sequential observations of a single item pair $Z^*_{ij}$ are sorted according to time, as in Table 3.4, deviations from a random walk indicate that the drift rate of this item pair is changing over time. However, many hypotheses can be tested. If for example the gender of test takers is available, one can sort the paired responses $Z^*_{ij}$ according to gender and test whether the observations are still random walks. Any sorting of Table 3.4 is allowed, whether it involves background variables of individuals, such as gender, or of schools such as school type, size, etc. In addition, not only categorical variables such as group membership can be evaluated, but also continuous background variables are easily sorted and evaluated, as long as the assumption of parameter invariance within each item pair or person pair is considered.

We propose a method to evaluate whether $Z^*_{ij}$ is a random walk for a specific pair of item $i$ and $j$. In Figure 3.1, the cumulative sums of a specific simulated item pair have been plotted\(^1\) on a time axis $t$. The item pair is simulated 1,000 times, and the 466 cumulative sums of the filtered $Z^*_{ij}$ are plotted, ignoring ties. Halfway through the simulation, a change in one of the item parameters is induced, making both items parallel, i.e., $\delta_i = \delta_j$. One can recognize the initial difference in the graph in item difficulties, causing an upward linear trend. Halfway through the simulation, this trend changes into a horizontal line as the difference in item difficulties disappears. We can regard this Markov chain as a Brownian bridge, i.e., the beginning and end points of the chain are fixed. Many methods are available to evaluate such a Brownian bridge, e.g., one can generate many Brownian bridges using identical beginning and end points.

\(^1\)This figure and all following figures are created using R (R Core Team, 2013).
points by permuting $Z_{ij}^*$ many times, and use these as references. Using the Kolmogorov distribution is another method to evaluate the extremes of a Brownian bridge. The 95% one-sample Kolmogorov bounds are used in Figure 3.1, indicating a large deviance from the reference distribution of the cumulative sums of $Z_{ij}^*$ of this specific item pair. Since the data is sorted according to time $t$, in applications such a deviation might be indicative of DIF over time of item $i$ or $j$, for which it can be useful to inspect related pairs.

Though a graphical inspection of an item pair can be useful to, for instance, detect what kind of a change occurs, one can easily apply a one-sample K-S test to evaluate the Brownian bridge. Since many pairs of items are often to be evaluated at the same time, we can for example construct a heat map of the significances of all item pair K-S test statistics. Item drift would show as lines in such a map. An illustration of such a map is given in Figure 3.2, which displays the result of a simple simulation of 10 items, where each item pair is answered 1,000 times. The item difficulties are drawn from a uniform distribution ranging from -2 to 2. For each item pair, all sequential pairwise updates were extracted, filtering out the $Z_{ij}^* = 0$ outcomes from Table 3.5. The number of remaining responses ranged between 30 and 468, with a mean of 219 responses per pair.

On the resulting Brownian bridge of each item pair, represented by a cell in the heat map, a K-S test was performed and its p-value plotted on the map.
Figure 3.2: Map of DIF between all item pairs.

A darker shade was used for $\alpha < .01$, a medium shade for $\alpha < .05$ and a light shade for the other p-values. Halfway through the simulation DIF was induced to a couple of items. Items H through J became easier by subtracting 0.5 from their item difficulties. For this set of items, the relative item difficulties with respect to all other items changed, which can be seen as both horizontal and vertical bands in the map in Figure 3.2 due to the symmetry of the matrix. Since the induced DIF is uniform for items H through J, there is no difference in difficulty between these items. Therefore, for the item pairs involving only these items, one can observe that there is no DIF in Figure 3.2. Also, the map shows that the number of observations used in this simulation using the specified effect size is not sufficient to produce significant differences for all cells. Though this simulation is just a simple illustration of how to use these non parametric tests to illustrate for example DIF over time, it can easily be extended, e.g., to illustrate differential item development in more complex training applications, as in section 3.3.3. Obviously, such heat maps can also be used to map person pairs, where idiosyncratic differences in development can be mapped and clustered.
3.3 Applications

3.3.1 Math garden

The methodology developed in this paper is applied to a dataset obtained from the Math Garden, a web based adaptive training environment to practice several mathematics abilities (Klinkenberg et al., 2011). The environment is mostly used by pupils from primary education, who can login either from school or from their homes, and practice basic math items. The Math Garden is adaptive in that it continually and concurrently estimates item difficulties and pupil abilities, and selects new items based on someone’s current ability estimate. An adaptation of the Elo rating system (ERS) (e.g., Elo, 1978; Batchelder & Ber-shad, 1979), including both response times (Maris & van der Maas, 2012) and parameter variance estimation (e.g., Klinkenberg et al., 2011; Glickman, 1999, 2001), is used to provide parameter estimates.

For the data examples hereafter data from the Math Garden domain of multiplication was selected. The item set was limited to include only the basic multiplication tables, e.g., question texts range from $1 \times 1$ and $2 \times 6$ to $10 \times 10$, 100 items in total. Data from the beginning of the school year of 2009 through the end of 2013 was used. The anonymized data was structured in a long format, like Table 3.4, cumulating to a total of over 6.3 million responses given by about 75 thousand pupils. Included are the background variables grade, gender, and an identifier for school membership. The number of responses per question ranged from just over 37 thousand to almost 80 thousand. The number of responses given by each pupil is quite skewed, with a median of 37 items and a mean of almost 84 items. Since we like to follow a cohort of pupils over time, we select all pupils with birth year 2004, which results in 1.4 million responses given by about 11 thousand pupils.

We applied the extended data filter in Table 3.6 in addition to the regular data filter in Table 3.5, where we assume parameter invariance within a single day.

3.3.2 Cumulative sums of a single item pair

We first focus on a single item pair, namely the symmetric items $2 \times 10$ and $10 \times 2$. Of the total number of 31,444 times that these two items occur in the data, 212 pairs remained after applying our regular filter and another 1,951 pairs were found in applying our extended filter, over 9 times more than the regular filter. To obtain these 2,163 item-pairs, the filters used a total of 8,228 responses from 2,285 persons. Clearly, the number of item-pairs available in
the data depend on the matchmaker, i.e., the item selection algorithm, that is implemented. Therefore, the number of available item-pairs also depend on the difference in estimated item difficulties used by the matchmaker. The cumulative sums of these item-pairs are displayed in Figure 3.3. Here, the cumulative sums of $Z_{ij}^*$ are plotted as a solid line. The plot is similar to Figure 3.1, but since we use observed time on the horizontal axis the figure is locally stretched or condensed depending on the distribution of observations over time. The result is that the linear one-sample Kolmogorov 95% confidence bounds therefore appear non-linear. To illustrate the density of the data, points are added to the graph for each day with data, and the amount of data points per day is illustrated by a histogram positioned at the bottom of the curve. The maximum number of observations for this specific item pair on a single day is 19.

The histogram in Figure 3.3 shows that the data is most dense after the year 2011. The graph indicates that there is a significant difference in the development of the two items, where the pattern seems that first $2 \times 10$ is easier than $10 \times 2$, after which they become parallel from 2012 onwards. We can split the random walk into two Brownian bridges, one prior to 2012, and one thereafter, which can be found in Figure 3.4 and Figure 3.5. Here we can observe that there is no significant deviance found for the chain after 2012, and indeed the
items are almost parallel since the relative probability correct is found to be 0.49, compared to 0.44 before the split. Such a development in item difficulties is of interest since a cohort is followed, and apparently these two items became just as difficult for these pupils.

We can apply the same idea to other item pairs, such as 1 x 10 and 10 x 1. Out of the 31,317 times that one of these items is observed, the amount of item pairs obtained using the regular filter is 240, and 2,179 item-pair observations are added using the extended filter, again over 9 times as many. A total of 9,196 responses remain in the filtered data, generated by 4,838 persons. The cumulative sums of this item-pair for the entire time frame can be found in Figure 3.6. Again, we split the time frame to obtain a figure for the item pair 1 x 10 and 10 x 1 before 2012 in Figure 3.7 and one thereafter in Figure 3.8. Again, we can observe how the item pair becomes parallel in Figure 3.8, with a relative proportion correct of 0.49 is obtained, compared to 0.56 before the split. It is interesting to observe that the direction of the linear trend of this item pair before 2012 is different than for the item pair in Figure 3.4. Though both item pairs become parallel after 2012, 1 x 10 starts out to be easier than 10 x 1, while 2 x 10 starts out to be more difficult than 10 x 2.

We like to add that some of the effects might be due to population artifacts, or population DIF. Since the system is driven by an adaptive algorithm, the most able pupils reach specific items at an early age, while less able pupils possibly reach these later. It is therefore possible that looking at e.g., Figure 3.6,
the subset of pupils from the 2004 cohort is quite different throughout time. If this is indeed the case, and DIF exists between different ability groups in this cohort, than population DIF and item DIF through time can be entangled.

### 3.3.3 Cumulative sums for multiple item pairs

For any subset or sorting of the dataset, it is possible to draw a map as in figure 3.2. For the cohort described in section 3.3.1, one sample K-S tests are performed on the cumulative sums for a number of item pairs. We chose all the item pairs of the multiplication tables of 9 and 10. Since the data is gathered using an adaptive system, the matchmaker determines the differences of item pair observations. These depend, amongst others, on the estimated item difficulties, see Klinkenberg et al. (2011) or Jansen et al. (2013) for a discussion on item selection in the Math Garden. The power of the tests is therefore in principle not equal for each item pair. However, using the extended filter in section 3.2.1 increased the number of responses sufficiently. All item pairs are observed, the number of observations on ranged from 725 to 8,821, with a rounded mean of 3,230 observations and a median of 2,896.

We created a DIF map in Figure 3.9, where each colored square indicates the test result of a single item pair. Three different shades are used, where the darkest indicates a significant test statistic at $\alpha < .01$, a lighter shade $\alpha < .05$ and
3.3. Applications

Figure 3.6: Cumulative sums of the item pair $1 \times 10$ and $10 \times 1$.

the lightest shade no significant deviation. White is used if no observations are present, which is only the case here for the diagonal. The figure is symmetrical around the diagonal, e.g., the order of items appearing within an item pair is disregarded. In general, one might approach such a heat map to search for patterns in the matrix in an exploratory fashion, e.g., using some sort of automatic sorting of rows and columns, or by applying data reduction methods. Another approach might be to formulate specific hypotheses concerning development of items and learning. Though we leave the interpretation of the heat map in Figure 3.9 to experts in learning basic arithmetic, we do observe several items such as $2 \times 9$, $10 \times 9$ and $10 \times 10$ that show differential development with respect to all other items in the map, and remarkably items such as $4 \times 9$ and $5 \times 9$ that barely show any differential development with the multiplication table of 10, but quite some differential development with the multiplication table of 9. We hope that presenting results in such a way can be instrumental in aiding discussion, especially since such maps can be made for different cohorts, years, for specifically boys or girls, for different school types, educational programs, and so forth.

In addition to inspecting item pair differences in item difficulty, the same method can be used to inspect person pair differences in ability. For example, differential growth between individuals can be used to detect drift with respect to reference groups, as well as cheating, population drift, differences induced by instruction method, etc.
3.4 Discussion

In this writing, we have described (extended) data filters to extract relevant information from large scale longitudinal data sets, such as usually generated by CAP or CAL environments and CATs. We regard relevant information as data that is informative of the relative difference in item difficulties and person abilities, regardless of their actual values. The approach as presented is non-parametric in that no parameter estimation takes place and simple cumulative sums on the filtered data, i.e., Markov chains, are used in conjunction with distribution tests, such as the one sample K-S test. Though simple, these filters allow to chart differential growth between pairs of persons and items, which can be quite challenging to model using statistical inferential methods using maximum likelihood estimation due to two reasons. First, the idiosyncratic developments of item and persons are influenced by many actors, including the estimations themselves due to reporting and feedback, which might result in a complex dynamic system. Second, the amount and sparseness of the data are typically challenging to the feasibility of methods using maximum likelihood estimation.

The data filters and their uses as presented here can be extended in many ways. Though we have so far ignored ties as being uninformative on the relative position of the persons or items in a pair, they do contain parameter dependent information and can be used in the construction of extended fil-
Figure 3.8: Cumulative sums of the item pair $1 \times 10$ and $10 \times 1$ from 2012 onwards.

ters, e.g., filters using two-parameter logistic model (2PL) models. Another extension is the use of the filtered data for parameter estimation, e.g., using the ERS on filtered data. For this specific context, the use filtered data solves the problem of drift in the rating pool.

An interesting use and extension of the methodology developed here, is the possibility to map a subset of persons or items that exhibit differential development as a set, in comparison to the entire set. The identification of such sets of items or persons is valuable from a diagnostic perspective, and might be especially useful in adapting education such as classroom instruction or policies. For example, groups of pupils with accelerated or stagnating development with respect to the group might be offered specific instruction, or items from specific item booklets that show differential development might be suspect for fraud. Also, the effect of a national curriculum can be evaluated over time by studying the differential development over time of the relevant sets of items, just as well as the effectiveness of instructional methods or programs. Many methods for the identification of sets of items or persons exist, of which the visual inspection of heat maps of K-S test results has been discussed already, possibly aided by manual or automatic sorting. Obviously, reduction or classification techniques such as multidimensional scaling, principal component analysis, k-means clustering, and many more can be used on these results to extract relevant subsets.
We like to add a note on the use of the RM in the aforementioned filters. Though the use of RMs can be considered quite restrictive in educational measurement, there are some considerations that ease some of these restrictions. First, though a RM is assumed to hold in applying the filter, it involves only the items and persons under consideration in any row of Table 3.5 or 3.6. Therefore, the RM parameters should be constant only during this observation, and are allowed to change before and after. In other words, just any RM is allowed to hold during a specified time frame, after which any other RM can hold, allowing in principle for quite some flexibility in applying such filters. Second, for an optimal CAT, the RM is actually an over-parametrized model (Zwitser & Maris, 2013) and therefore possibly not so restrictive for the data under consideration. However, extensions in using a more general item response models, e.g., the 2PL, can be useful in some data configurations. The appearance of ties is informative in this situation, but only preliminary research has been conducted in

![Map of DIF over time between item pairs involving multiplication tables 9 and 10.](image-url)
this direction.

A curious finding is the similarity between data filters described in this writing and data filters used in the context of generating so-called unbiased bits from unequal pairs of biased bits in the field of random number generation (von Neumann, 1951; Peres, 1992). Conceptually this relation is interesting in that in both applications, the filters are used to remove bias from single observations by considering pairs.

A final note concerns the possibility to change the sorting of the data to test for different hypothesis. Though the context that is elaborated upon in this writing is differential development of pairs of items over time, obviously the same methodology hold for individuals and for different ways of sorting the data. Testing for differential development through time is just as easy as testing for differential development of genders, of different school types, of item types, of population subgroups, etc., which we believe to be a quite powerful feature in the use of such filters.

References


References


