ADAPTIVE Matrigma®

TECHNICAL MANUAL

2017 EDITION

HUNTER MABON
FREDRIK NIEMELÄ
ANDERS SJÖBERG
SOFIA SJÖBERG

ASSESSIO

Editing and psychometrics: Fredrik Niemelä, Anders Sjöberg and Sofia Sjöberg

Graphic design: Lina Strand

Content

Introduction

This manual describes the background and development of Adaptive Matrigma, Version 1.0. Adaptive Matrigma is a further development of the non-verbal general mental ability (GMA) test, Matrigma[1]. Because of the connection between these tools it is recommended that readers familiarize themselves with the technical manual for Matrigma (Mabon & Sjöberg, 2016) before reading this manual. The Matrigma technical manual presents, for example, the theoretical background, the importance of GMA in general, the measurement of this ability in particular, the relevance of matrices as a format, and the connection with outcomes such as performance at work. These areas are also relevant to Adaptive Matrigma. Note also that this manual presumes the reader has a basic knowledge of psychometrics and is well acquainted with Classical Test Theory. For those who need a refresher on these subjects, the books Scale Construction and Psychometrics for Social and Personality Psychology (Farr, 2011) and (in Swedish) Arbetspsykologisk testning (Mabon, 2014) are recommended.

This manual begins with an account of the rationale behind the development of Adaptive Matrigma, before presenting the underlying measurement and psychometric models and how they differ from those of Matrigma. This is followed by a description of the development of Adaptive Matrigma and a report of the results from the initial evaluation and psychometric analyses. The manual concludes with a discussion of the recommended areas of application along with recommendations and guidelines for administration, utilization, and the interpretation of test results.

The need for the Adaptive Matrigma

The demand for tests measuring GMA for use in work life psychology has increased immensely over the last fifteen years, and the amount of testing has increased dramatically. This is generally positive, as it implies that there is greater competence in the area of personnel selection within companies and organizations, and that more rigorous and relevant individual assessments are being conducted. This increased use also leads to higher demands being placed on the tools and, in turn, a greater call for continued development of assessments. Matrigma was published for the first time in 2008, at which time GMA testing in work life situations was still relatively uncommon, although interest was beginning to gain momentum. Matrigma quickly became popular and the number of test administrations has been increasing at a steady rate every year since then. Matrigma has been undergoing continual improvements, with the development of new matrices and response alternatives as well as revised versions. A number of new parallel versions have also been introduced, and Matrigma currently includes five parallel versions (Mabon & Sjöberg, 2016). Matrigma is now a well-established and renowned test within work psychology; it is available in over twenty

[1] The product name 'Matrigma' refers to the product comprising five fixed parallel versions and is the product meant when the 'classic' version of Matrigma is referred to. The product name 'Adaptive Matrigma' refers to the product comprising an adaptive version based on a 3-parameter IRT model.

languages and used across the world.

In 2016, nearly 100,000 test administrations using Matrigma were carried out. In recent years, the need for a new tool using matrices as stimuli and with a shorter administration time has increased, as the possibilities of developing such a tool for practical use in individual assessments within working life have also increased. These needs and possibilities can be summarized in the following points:

Increased mobility in the labor market implies an increased need for conducting more individual assessments in association with recruitment and selection

- Increased use of testing in general and GMA tests for selection purposes in particular
- Increased demand for selection tools requiring less time from the candidates; this would be a benefit to candidates (who in general nowadays apply for more jobs during a lifetime compared to previous generations) *and* to test administrators/organizations (who often compete for candidates who also apply for work elsewhere and therefore need flexible and easily accessed tools)
- Large testing volumes mean increased exposure, which makes it difficult for the test provider to maintain the intellectual integrity of assessments over time – even with the use of parallel versions or item[2] banking[3] solutions
- Research in the area and continuous method development have helped produce methods that are useful in practice and more precise for achieving this type of level determination in a shorter time
- Advancements in the digital sphere have made it possible to implement the models that are required to be able to deliver high quality tools for the intended purposes in a cost-effective way

These points summarize the rationale behind the decision to develop Adaptive Matrigma.

Assumptions

Matrigma and Adaptive Matrigma, as mentioned above, have a good deal in common. They share a theoretical basis with regard to the history of intelligence (GMA), Spearman's *g* factor, and their approach to measuring this characteristic. The tests have similar purposes, areas of use, and connections with relevant outcomes such as job performance. They also share a matrix format for stimuli and they do in fact share items (tasks); both tools use the same set of items. This implies that *both tools estimate an*

[2]The word 'item' refers to a unit and can, in psychological testing, be represented by, for example, an exercise, a task, or a statement that a test person is to take a position on. In Matrigma and Adaptive Matrigma, an item is comprised of problem-solving tasks in the form of a matrix with six response alternatives. In this manual the terms 'item' and 'task' are used interchangeably.

[3] Item banking traditionally refers to a database of items, often classified into groups based on certain characteristics, from which they are selected for use in, for example, a testing administration.

individual's probable level of GMA, which is in turn a good predictor of future job performance. How they each achieve this differs, however. The differences at the root of this are rather abstract and technically advanced, but they do have an impact on administration, testing experience, and the calculation of results – although the results themselves carry the same meaning and are to be interpreted in a similar way. The differences are fundamentally due to the fact that Matrigma and Adaptive Matrigma are based on different *measurement theories*: Classical Test Theory (CTT) and Item Response Theory (IRT) respectively. Since they are based on different principles and assumptions, the psychometric models, and thus the standards for defining aspects such as validity, reliability, measurement error and so forth, differ as well. In order to understand how IRT-based Adaptive Matrigma works, it needs to be considered in relation to Matrigma and CTT.

Classical Test Theory

CTT has always been and still is the most common basis for method development. Most psychological tests, and other measures, rely on this theory even if it is not always explicitly stated or even known to users. CTT is built on a predetermined set of items that are predefined (a priori) within a group-level conceptual framework, and the items included in the test are presumed to be at the interval scale level (in that the data are ranked and that the distance between the scale points are presumed equal). In other words, the item set is predefined and constant even if there are an endless number of variations of these items and the fixed set is a random selection of them. Since CTT is based on correlations, and the final results (often the total number of correct responses) are presumed to rest on interval data, it is just as difficult, for example, to increase one's performance by one point (1 correct response) across the entire scale.

According to CTT, every response that an individual gives is presumed to consist of a true value plus a certain amount of error. The measurement errors are presumed to negate each other when the items are summarized into a scale, as these errors are presumed to be random in CTT. This entails that it is impossible to obtain a true score for a characteristic as the measurement will always contain a certain amount of measurement error. This random error variation is defined within CTT as reliability and is always estimated at the group level (thus not at the single individual test score level). Measurement error on the group level, however, may be used to estimate the error in a single individual test score – known as the standard error of measurement. Since the interval scale level is presumed in the measurement, the recognized features of the normal distribution are also presumed, implying that the standard error of measurement is equal across the scale.

According to CTT, the determination of level (in this case GMA level) for an individual is made in relation to a predetermined level that is based on a distribution of test scores. This distribution of test scores defines the norm group. The distribution of the norm group is of great importance and it is crucial that there is variance in the test scores (among the individuals).

Item Response Theory

In the 1950's, development of the model now known as Item Response Theory (IRT) began, and in the 1970's it started to receive increased attention, especially from those interested in method development. The complete IRT model utilizes three parameters to indicate the characteristics of *individual items*: difficulty (*b*), discrimination (*a*), and guessing (*c*) (DeMars, 2010).

The first parameter, difficulty (*b*), represents the level of difficulty for each item, and it is defined on the same scale as the underlying characteristic (for Adaptive Matrigma; GMA). The characteristic is arbitrary, or latent, but it is usually anchored so that the distribution of the characteristic in a given group has a mean value of 0 and variance of 1. Difficulty (*b*) thus identifies the level at which 50% of those tested are expected to respond correctly to the item. In CTT, difficulty is labeled *P* and defined as the proportion of individuals (in the norm group) who responded correctly to an item; difficulty corresponds to the mean value for each item.

The second parameter, discrimination (*a*), represents the extent to which an item differentiates between individuals who are at different levels for a given characteristic. Discrimination (*a*) is therefore a measure of the effectiveness of an item, with high discrimination (*a*) being desirable. This parameter is sometimes called 'the slope' since an item's characteristics according to IRT are often illustrated using an Item Characteristic Curve (ICC) and the degree of discrimination is illustrated by the slope of the curve. The point-biserial correlation coefficient is the typical measure of discrimination within CTT. A positive correlation indicates that individuals who responded correctly to an item have a higher aggregate score for the remaining items compared to those who responded incorrectly.

The third parameter, guessing (*c*), represents the likelihood of responding correctly to an item by chance or through guessing. This parameter is primarily determined by the number of response alternatives: five response alternatives, for example, yields a 0.2 likelihood of guessing correctly. The likelihood of guessing correctly, however, is also influenced by the level of difficulty for each response alternative. For example, if two out of five response alternatives are obviously incorrect, the guessing is de facto among the three remaining alternatives, yielding a 0.33 likelihood of guessing correctly. By correcting for guessing (*c*), more exact estimates of an individual's level can be determined.

In most cases, one or two parameters are utilized, with guessing and discrimination being set to 0 in the former and guessing set to 0 in the latter. This usually occurs during method (test) development. For example, the 1-parameter model is often utilized to determine the degree of difficulty of items in order to rank and administer them according to increased level of difficulty. Such a model was used, for example, to rank items in the development of Matrigma's parallel versions (Mabon & Sjöberg, 2016). Advancements within information technology (IT) in recent decades have made it possible to utilize 1-, 2-, and 3-parameter IRT models for *administration* purposes, when selecting what items are to be administered in each administration (based on the response to the previous item), and for *scoring* the results. This type of testing is called computer adaptive testing (CAT).

Advantages of IRT

Using IRT-based CAT with all three parameters has many advantages. Items that differentiate better, or more reliably, are utilized more effectively. This implies that the final test scores are assumed to be more reliable in comparison with test scores that are derived from CTT, where all items are treated equally. It also implies that if different individuals are administered different versions, the test scores will be adjusted in accordance with the degree of difficulty. Another advantage is that the items that an individual is administered are selected according to his or her level on the characteristic being measured. This should better maintain the respondent's interest and avoid frustration by minimizing the number of overly easy items and excluding overly difficult items. However, such adaptive testing may also be perceived as challenging for respondents since it quickly fixates on identifying an individual's maximum level of performance. How this is handled in Adaptive Matrigma (with a warm-up phase) is described below. With IRT-based CAT it is possible to administer an item according to its ability to discriminate (*a*) and in accordance with its guessing value (*c*) for each response alternative. Thus, the next item to be administered may be adapted depending on the specific response to the previous item. The model then plots the total scores on the same scale, making them comparable; this is not possible with CTT. Lastly, by presuming that there is no measurement error for individual items and that the items are independent of each other, it becomes possible to estimate the precision, or reliability, of an *individual's* specific test score.

In contrast to CTT, IRT provides information about individual differences by determining the characteristics of single items *together* with the individual's responses. This enables a test administration to adapt according to each individual's specific response to each specific item. The fact is that IRT is not really a measurement theory since the model does not presume that there is a true value or measurement error (Kehoe & Sackett, 2010). The IRT model, in contrast to CTT, is independent of the sample distribution (norm group within CTT) that is the basis for the estimate. This allows for the results to be interpreted in a different way than with CTT. Since IRT compares items and individuals on the same scale, an individual's position on the scale can be identified, independent of how others have responded on the same scale. When the parameter estimates (difficulty (*b*), discrimination (*a*), and guessing (*c*)) for all of the items have been determined, the items can be placed on the same scale. The common, overlapping items make it possible to develop a large number of items with known measurement characteristics. All of the measurements then express the results on the same exact scale. This means, for example, that IRT does not require that two measurements are parallel in order to obtain comparable results, something that is required with CTT.

It should be noted that IRT is based on the assumption of unidimensionality, which holds that all items measure the same underlying construct (characteristic, in Adaptive Matrigma; GMA), and on the assumption of local independence, whereby items are considered to be entirely independent of each other. But even though this is very difficult, if not impossible, to achieve in practice, IRT has shown to be a useful approach to solving many practical measurement issues even when all of the assumptions are not entirely met. Another example is the assumption of interval scale levels in the data. Since IRT does not rely on correlations, it is not required that response alternatives be on interval scale levels; with IRT, data may actually be combined on all types of scale levels (interval scale,

nominal scale, and ordinal scale). For readers more interested in the IRT approach, the appendix provides some suggested further readings on the topic.

Development

The development of the Adaptive Matrigma occurred in two main steps. In the first step a pilot version was developed using data from candidates (N=1000) who had participated in one of the five fixed parallel test versions that were produced in 2014 for the classic version of Matrigma. A total of 150 items were analyzed using a full 3-parameter IRT model. The items and the three estimated parameters for each item were implemented in the web-based platform Ascend by Assessio in order to calculate the results for the pilot version of Adaptive Matrigma. The pilot version was implemented along with candidate instructions and practice tasks primarily based on the corresponding content for the classic version of Matrigma.

Results based on IRT-models are expressed as a so-called theta value, denoted by θ. As can be concluded from the above, a theta value is not calculated based on a comparison against a distribution of test scores, as CTT postulates with its norm group. Thus, information regarding an individual's level in comparison with others (referring to a distribution, a norm group) is not generated through an IRT-based model. The theta value, which represents an interaction between the difficulty level of an item *and* the individual's level may for Adaptive Matrigma vary between –3 and +3. The theta value is thus suitable for ranking the candidates on the characteristic being measured (GMA when using Adaptive Matrigma) but does not indicate an individual's relative level on the characteristic in relation to a distribution, a norm group.

In the second step, in order for the test to be able to generate even more exact results, a study was conducted using the data that had been collected on the pilot version. The aim of the study was to obtain more refined parameter estimates. This study utilized data from candidates who had taken both Matrigma *and* Adaptive Matrigma for selection purposes (N=1,667). Everyone in this sample had been administered one of the five parallel versions of Matrigma and received a C-score[4] indicating their GMA level in relation to the applied norm group of N=4,606. The data analysis commenced with estimating new and refined parameters for Adaptive Matrigma based on the group of N=1,667. Employing the new parameter estimates for all items, a new theta value was then calculated for each candidate. The group's (N=1,667) theta values thus formed a distribution which could be linked to the distribution of C-scores that the same candidates had obtained from the classic (norm group based comparison) version of Matrigma. In this way, the theta value distribution for Adaptive Matrigma could be 'translated' into a C-score distribution.

The theta-based C-score obtained through Adaptive Matrigma thus provides information about both the absolute level of GMA *and* the relative level – in comparison to the Matrigma norm group of N=4,606 (see the Matrigma Technical Manual for more

[4] A C-score is a standardized score based on the C-scale which ranges between 0 and 10, has a mean of 5, and a standard deviation of 2.

information about the norm group).

Overall, the result of implementing the updated new parameter estimates will be to generate a somewhat higher C-score (approximately 1 C-score) compared to the previous pilot version. Note also that when implementing the new parameter estimates, the instructions and practice tasks were also refined and updated, affecting the candidates' experience. In general, by applying a refined user experience approach, the length of text based instructions were shortened and simplified and the practice tasks were made more instructive.

Principles for administrating the items and calculating the results

The IRT model provides the framework for how Adaptive Matrigma functions at a general level. Implementing this type of model for practical use, however, requires making decisions on a number of aspects that are not specified or regulated in the model. Sometimes it is also, for practical reasons, necessary to deviate from the model, and it is important for the test administrator to be aware of the deviations as well as the rationale behind them. The following therefore gives a description of the factors that influence which items are selected for each administration and clarifies how the choice of items are made in several steps. But first a summary and a reminder of the differences between CTT and IRT regarding the calculation of results is presented.

As described above, applying CTT implies that the estimation of an individual's level (of GMA or some other characteristic) is usually made by calculating the number of correct responses, giving one point for each correct response to an item. The number of correct responses are then tallied to create a sum score. The sum score is then compared to the distribution of sum scores (provided by the norm group). When applying IRT, the individual's level is instead calculated from an iterative process in which the parameters of difficulty (*b*), discrimination (*a*), and guessing (*c*) are weighted together along with each individual's response to each item. The combined level is summed into theta (θ). It is possible to express theta on any suitable scale, but Adaptive Matrigma, like Matrigma, uses the C-scale (Mabon & Sjöberg, 2016).

A key detail about the use of IRT is that, strictly speaking, the test should always begin with administering a task that is at the average difficulty level, that is, at a C-score level of 5. A task is considered to be of average difficulty when respondents who are at an average level regarding the characteristic in question have a 50% likelihood of answering correctly (or incorrectly). The principles behind which tasks are to be administered at the beginning of the Adaptive Matrigma deviate from this principle. The main reason for this is that if the beginning is too difficult, there is a risk that it could negatively affect the rest of the testing experience and, in turn, the final test results. The Adaptive Matrigma always starts with what could be called a warm-up phase. It should be noted that these tasks are being scored and are separate from the practice tasks that respondents go through before the testing starts.

To determine an individual's level of GMA the following steps are taken:

1. A random item with known characteristics for the parameters of difficulty (*b*), discrimination (*a*), and guessing (*c*) is administered to the respondent. The first item, item 1, to be administered is always a randomly selected item whose difficulty level corresponds to a C-score of approximately 1.

2. The respondent's correct or incorrect response to the item is registered.

3. If the response to the first item was *correct*, the second item will be a randomly selected item from among those corresponding to approximately a C-score level of 3, a somewhat more difficult item than the previous. However, if the response was incorrect for item 1, item 2 will be a randomly selected item from among those with a difficulty level corresponding to somewhat lower than a C-score of 3. This means that even if the respondent gives an incorrect response to item 1, the next item will be somewhat more difficult – although not as difficult as it would have been if the response to item 1 had been correct.

4. The selection of item 3 is based on the same premises as the previous item except that the randomly selected item following a *correct* response will correspond to a difficulty level of approximately a C-score of 5, a somewhat greater difficulty level. An *incorrect* response to item 2, however, is followed by a randomly selected item that corresponds to a considerably lower difficulty level. A respondent can thus deviate from the 'warm-up phase' earlier than another respondent.

5. This is based on the fact that the initially selected items are on the low side in terms of difficulty level. Consideration is given to the responses to the previous items but the selection of the next item is on the low side in terms of difficulty level. The items that are administered in this warm-up phase can therefore be considered adaptive, but to a somewhat limited extent. This assigning of relatively lower difficulty decreases successively as the matching gradually improves along with further items being administered.

6. When the 'warm-up' phase is completed, the respondent is assigned an initial theta (θ) value called the 'prior.' The prior is decisive in the next step of this iterative process.

7. In the next step, item selection is based on previous responses and the values of the three parameters for a larger group of individuals with the purpose of "finding" the respondent's "true" θ value. The difference between the respondent's θ value and the likelihood that the person will respond correctly to the next item is registered. The sum of these standardized differences is added to θ, with negative differences decreasing the θ value and positive differences increasing it. The best estimation of a respondent's GMA level is then considered to be reached when an item response minimally alters the respondent's θ value.

8. This iterative process continues until two minutes (out of the 12 minutes in total) of test administration time remain. For the last two minutes, the items are no longer selected randomly among those at a certain difficulty level. They are

instead selected according to the characteristic of their *maximum smallest distance in difficulty* to any other already answered task. The purpose is to probe the difficulty level as effectively as possible and decrease the possibility for substantial selection bias of similar items. The latter would be disadvantageous since this data may end up as the basis for future re-estimations of the parameters and may affect the respondents' experience negatively, which is described in the following.

Perceived item similarity by respondents

In connection with the trial testing of the pilot version it was found that respondents at times felt that the administered matrices were very similar. This gave rise to a sense of monotony and also raised a number of questions both during and after the testing. In response to this, item selection also takes the "families" of items into consideration. The families are comprised of tasks that are *perceived* as identical or very similar by the *candidates*. The purpose of taking this into account is to provide the candidates with a testing experience that is perceived as more varied and, above all, to avoid questions and concerns being raised over whether the "same task" was administered several times. This may demand more time from the candidate and affect the final results. The tasks are therefore randomly selected from among 1) tasks at the same difficulty level but (2) which were not within the same family as any of the ten previous items.

Time limitations and the number of tasks

According to a strict utilization of an adaptive IRT model, the number of tasks could theoretically range from a single task to any infinite number of tasks. For practical reasons, however, an administration of Adaptive Matrigma takes less than 12 minutes for a respondent.

In addition, each item is given a one minute time limit. This limit is based on the requirements that the test items should be quick to administer but at the same time generate test scores that are of high quality in terms of validity and precision. A certain number of tasks need to be responded to in order to be able to generate reliable results. It may be hypothesized that being under considerable time pressure would affect the degree of difficulty of an item. However, whether this effect actually exists and if it has any effect on the accuracy of test scores remains to be investigated.

Besides the 12-minute timeframe, there is a limit of 40 tasks. Regardless of the responses given, a respondent will never be administered more than 40 tasks within the 12 minute testing time. This limit is set in order to maintain the intellectual integrity of the tasks. Furthermore, it is considered extremely unlikely that a respondent could respond to more than this many items in a genuine and serious manner within 12 minutes. Whether this limit should be further restricted will be shown by further research.

Instructions for use and interpretation

Areas of use

Adaptive Matrigma has been developed for assessment in the context of personnel selection. The qualities measured are universally important and impact job performance in all professions, therefore making Adaptive Matrigma applicable for any position and for all industries and businesses. Adaptive Matrigma may be used for professions at all levels and in all lines of work, preferably as a first step in the selection process. Adaptive Matrigma is not intended to be used in a development context such as in manager and employee development, career guidance, team building, coaching etc., or for use within a clinical context.

Note that the format of delivery (web-based and unsupervised administration) and the time efficiency of Adaptive Matrigma makes it highly suitable for screening a larger number of candidates. However, there is no inherent barrier for using Adaptive Matrigma on a smaller number of candidates.

Administration and scoring

Adaptive Matrigma is available via Ascend by Assessio and through partner systems via Ascend's API[5] functionality. The respondent completes the items shown on screen and the web system computes the raw scores, converts the raw scores into standardized scores, generates results, and provides standardized feedback reports. The use of Adaptive Matrigma requires a trained test administrator who may choose to either administer Adaptive Matrigma remotely by sending a link to the respondent via e-mail, or to administer Adaptive Matrigma supervised on-site. It is recommended that Adaptive Matrigma be administered under supervised conditions. If a respondent completed Adaptive Matrigma unsupervised, it is recommended that the respondent be re-tested under supervised conditions or that the test score is supplemented with results from an additional GMA test.

Requirements for testing

The requirements for administration and conditions of testing are:

- A 12-minute timeframe for responding to Adaptive Matrigma. The time limit is tracked by the web system and is shown on screen; when the time limit has been reached, the test session will end and the respondent's responses will be saved.
- Each item has a time limit of 1 minute. The time limit is tracked by the web system and is also shown on screen; when the time limit has been reached, a new item will be administered and any response selected by the respondent will be saved.

In order for the respondent to perform at maximum capacity and to experience a fair,

[5] 'API' is an abbreviation of Application Programming Interface and helps companies to share data in a controlled manner.

reliable and valid assessment, is responsible for:

- Ensuring respondents' basic reading comprehension – although the written instructions aim to be short, simple, and straightforward, they nevertheless require a basic level of reading comprehension.

- Ensuring that the respondent does not suffer from any form of impairment that is likely to have a negative effect on the test result. This may include but is not limited to perceptual, visual, and/or cognitive impairments.

- Ensuring that the Adaptive Matrigma is responded to in a non-distracting environment – public environments, such as internet cafés and public transportation are not suitable for taking Adaptive Matrigma.

- Ensuring that respondents' access Adaptive Matrigma in the most suitable way. It is recommended that the test is completed using a personal computer with a full-sized computer screen as Adaptive Matrigma has been visually adapted and developed for such administration conditions. The technical information reported in this manual is based on assessments conducted under such circumstances, implying that the quality of the assessment apply only to such conditions. Note that it is possible for the test administrator to provide respondents' with access to Adaptive Matrigma using a mobile device, such as a smartphone or a tablet. Responding to Adaptive Matrigma on such a device may however affect the result. This may have implications for interpretation of a single score since the technical information is based on administrations using a personal computer. It may also have implications for comparisons *between* scores, especially when different devices have been used. It is the test administrators' responsibility to inform and ensure that respondents' access Adaptive Matrigma in the preferred and appropriate way. If the test administrator has allowed for the use of multiple devices this will be shown in Ascend with an icon symbolizing the mode of device use by each respondent.

- Ensuring that the respondent has access to a personal computer when taking the Adaptive Matrigma – it is not recommended to use a tablet, smartphone or similar device as Adaptive Matrigma has been visually adapted and developed for administration on a full-sized computer screen. Using other devices may affect the test result.

- Ensuring that the respondent has basic technical skills – the respondents' must for example be able to use a mouse and/or keyboard in order to complete Adaptive Matrigma. The test administrator should ensure that the technical aspects do not increase the test difficulty for the respondent, as this would have a negative effect on the result.

- Ensuring that the respondents' has access to a stable and reliable internet connection for the full duration of the testing in order to ensure a valid result.

Information for respondents before testing

If Adaptive Matrigma is to be administered unsupervised, thus remotely, the test administrator sets this up through the web system. The test administrator will require the e-mail address of the respondent. The test administrator will be provided with an e-mail template containing a link to the test and some basic information. This e-mail is editable; the test administrator may thus insert specific information for a single respondent or a group of respondents. It is strongly recommended that the e-mail to the candidate include information regarding:

1. The purpose of the testing.
2. What type of test Adaptive Matrigma is and why it is being used in the present context.
3. How Adaptive Matrigma will be administered and what is required for completing the test (see Requirements for testing in this section).
4. What mode of device is to be used. This is especially important if the respondents' are given access to complete Adaptive Matrigma using mobile devices.
5. How the results will be used and saved, by whom and for how long.
6. The respondent's right to choose whether the test score should be used as part of the information provided about him or herself for the selection process.
7. If feedback will be provided; if so, when will it be distributed, what format will it be in (standardized on screen, personal feedback face-to-face, over the phone), and what will it contain.
8. Contact details for the test administrator.

More information about the rights and obligations of test distributors, test administrators, and candidates are to be found in international guidelines for testing (e.g., www.intestcom.org, www.efpa.eu/professional-development, www.iso.org/standard/56436.html) and is often provided by national psychologists' associations.

Presentation and interpretation of results

The results are presented on what is known as the C-scale, a type of standard scale, in order to facilitate interpretation and comparison. As described above, the C-scale ranges from 0 to 10, has a mean of 5, and a standard deviation of 2. In order to facilitate the interpretation of test scores, the C-scale has been divided into three levels, representing low scores that are below average (0-2 C-scores), average scores (3-6 C-scores), and high scores that are above average (7-10 C-scores). Thus, the low scores that are below average correspond to approximately 16% of the lowest scores in the norm group; the average level corresponds to results that are plus/minus one (1) standard deviation from the mean of the norm group; and the high level, above average scores correspond to

approximately 16% of the highest scores in the norm group.

The characteristics measured by Adaptive Matrigma, abilities such as finding logical, sometimes hidden connections, conducting abstract reasoning, making logical conclusions, and solving novel problems, all vary between individuals and are important in a work context. In general, the higher a respondent score on Adaptive Matrigma, the more likely it is that he or she will exhibit good job performance. Conversely, the lower a person scores, the less likely it is that he or she will exhibit good job performance.

Standardized feedback reports

After the testing is completed, the web system will generate a standardized score, a C-score, and two types of result reports for each respondent: the Interpretive Report and the Your Result feedback report.

The standardized feedback report, labeled Interpretive Report, is in the form of a pdf document and intended for the test administrator. This report contains information about the respondent's C-score and level (defined as Low, Average, or High according to the above) along with a more in-depth account of what the results mean. This includes descriptive text regarding general mental ability, norm group comparison, and the meaning of the different levels.

The second standardized feedback report generated by the web system is labeled Your Result. This report is shown on screen to the respondent, if this is enabled by the test administrator. It is thus optional for the test administrator to provide the respondent with this feedback (set up in project management). The content of this report is considered to be self-explanatory and does not require personalized feedback. This report contains information about the respondent's level, expressed as Below average (labeled Low in the Interpretive Report), Average, or Above average (labeled High in the Interpretive Report), and a description of what the results mean. It also provides information about what Adaptive Matrigma measures, what the results mean regarding comparison against a norm group, and what to remember when reading the results.

In addition to the individual reports, the C-scores of all respondents tested within a project are listed on screen. The Ascend user interface also enables ranking of respondents based on their C-scores. The intention of the project overview is to provide a basis for decision-making at the group level.

References

DeMars, C. ((2019). *Item response theory. Understanding statistics measurement*. Oxford, New York: Oxford University Press.

Furr, R.M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. SAGE Publications. ISBN 9780857024046.

Mabon, H. (2014). *Arbetspsykologisk testning*. Assessio International: Stockholm, Sverige.

Mabon, H., & Sjöberg, A. (2016). *Matrigma. Technical Manual*. Stockholm: Assessio International.

Kehoe, J.F., & Sackett, P.R. (2010). Validity consideration in the design and implemantation of selection systems. In Farr, J.L., & Tippins, N.T (Eds.). *Handbook of employee selection* (pp. 56-92). New York, NY: Routledge.

Appendix

Suggested readings on Item Response Theory

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Psychology Press. ISBN 978-0-8058-2819-1.

Baker, F. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Baker, F. B., & Kim, S-H (2004). *Item Response Theory: Parameter Estimation Techniques* (2nd ed.). Marcel Dekker. ISBN 978-0-8247-5825-7.

van der Linden, W. J., & Hambleton, R. K., eds. (1996). *Handbook of Modern Item Response Theory*. New York, NY: Springer. ISBN 978-0-387-94661-0.

de Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer. ISBN 978-0-387-40275-8.

Fox, J-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer. ISBN 978-1-4419-0741-7.