THE CONCEPT OF DRIFT AND OPERATIONALIZATION OF ITS DETECTION IN SIMULATED DATA

by

Keren Roded

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate and Postdoctoral Studies

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

September 2017

© Keren Roded, 2017

Abstract

In this paper, the phenomenon of changes in item characteristics over time (often referred to as *drift*) is discussed from several theoretical perspectives, and a new procedure for the detection of Item Parameter Drift (IPD) is proposed. An initial evaluation of the utility of the proposed procedure is conducted using simulated data modeled by the 2-Parameter Logistic (2PL) Item Response Theory (IRT) model. In addition to the proposed procedure, an IPD analysis of the simulated data is conducted using two known methods: Kim, Cohen, and Park's (1995) extension of Lord's (1980) Chi-square test of Differential Item Functioning (DIF) to multiple groups, and logistic regression. The results indicate high agreement and accuracy in the detection of true IPD using the two known methods, but poor performance of the proposed procedure. Possible explanations of the findings and future directions are discussed.

Lay Summary

The use of tests and assessment tools over extensive time periods is a common practice in many fields and contexts. When the same question, or task, is presented to individuals at different points in time, its functionality in measuring the phenomenon of interest may undergo changes due to, for example, social processes or historical occurrences. Such changes, commonly referred to in the psychometric literature as Item Parameter Drift (IPD), are of central concern in any ongoing measurement framework. This paper offers some unique theoretical perspectives through which these changes can be conceptualized, and proposes a new procedure for evaluating IPD which can be applied using various measurement models. The paper also presents an initial application of the proposed procedure, and evaluates the utility of the application in relation to two existing methods using simulated data.

Preface

This thesis is the original work of the author, Keren Roded, under the supervision of Dr. Bruno Zumbo.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DIF	Differential Item Functioning
ICC	Item Characteristic Curve
IPD	Item Parameter Drift
IRT	Item Response Theory
LogR	Logistic Regression
MGDIF	Multiple-Groups DIF
2PL	2-Parameter Logistic
3PL	3-Parameter Logistic

Introduction

The primary purpose of this paper is to introduce an approach for detecting Item Parameter Drift (IPD) which, to the best of the author's knowledge, has not been presented in the psychometric literature so far. The central aim of the paper is to provide a proof of concept of this proposed procedure: to illustrate its use with simulated data, and to compare its conclusions with those of well-established methods for detecting violation of measurement invariance across multiple groups.

At the onset of this research endeavor, it seemed reasonable to assume that the proposed procedure should allow for identification of IPD in a manner which is (1) robust to differences between the ability distributions of test takers at various time points, (2) independent of the nature of changes over time (i.e., no specific pattern of drift is tested), (3) sensitive to small/non-significant changes between consecutive administrations which accumulate to larger/significant changes over time, and (4) applicable to tests in which different models are fitted to different items. This paper was motivated by the potential achievement of these promising advantages, which are not necessarily obtained by currently available methods for evaluating IPD.

In order to contextualize this demonstration, the paper begins with a discussion of several perspectives through which drift can be defined and conceptualized. The new procedure is then introduced, and simulated data are analyzed for IPD using this new approach along with two additional methods: logistic regression, and the procedure for detecting differential item functioning between multiple groups introduced by Kim, Cohen,

& Park (1995). Lastly, the outcomes of all three methods are compared in order to evaluate the utility of the proposed procedure.

Conceptualization

Three Dimensions of Test Stability

The process of test construction frequently begins with defining a construct of interest, developing tasks which elicit responses based on the defined construct, grouping tasks into test forms, and selecting a measurement model to be applied for assembling individual scores based on the task responses. At this point, two fundamental concepts of measurement are often addressed: test *reliability* and *validity*. In broad terms, reliability represents an inner characteristic of the measure referring to its internal consistency, or stability of the scores under repeated assessments of the same individuals. In even broader terms, the concept of validity represents the extent to which the test constitutes a reliable measurement of the desired construct, or the degree of justification for specific uses of the test results. That is, a test may be consistent and reliable, but those qualities are of limited utility when the broader aspects of its uses and interpretations are ignored. When a test is being used over considerable time periods, a third concept comes into play: the stability of the test characteristics over time, or lack of ***drift***. This concept expands the general view of test stability beyond reliability and validity in the sense that a measure may be highly reliable and possess strong evidence of validity of its inferences, but when the possibility of drift is ignored, those qualities are limited to the context of one specific point in time.

While efforts to enhance a measure's reliability and validity are often considered basic requirements for any testing framework, the third concept – lack of drift – is less commonly addressed, and investigations of stability over time are often centred around technical

methods for locating *Item Parameter Drift (IPD)*. Drift is often narrowed down to changes in item parameters over time, or even to changes solely in item difficulty. The fact that the term 'drift' defines a negative and undesired outcome of using test items over time, in contrast to the positively worded 'reliability' and 'validity', is indicative of the practice from which its investigation has emerged: drift is often regarded as a threat to be dismissed rather than "lack of drift" being considered a quality to strengthen. This distinction may be explained by the fact that the stability of test characteristics can not be assured solely based on the developer's efforts; changes in the characteristics of tests and measures may be observed due to events which are beyond control (e.g. historical occurrences). Therefore, a test can only be considered stable regarding its past administrations, and monitoring stability over time is an ongoing effort by definition. It should be noted that the same kind of ongoing practice is applicable for the task of validating tests and measures; therefore, given the large scope of common definitions for test validity, investigation of drift can be viewed as part of the validation process, which involves a collection of evidence in order to support and justify the use of tests for specific purposes over time.

From a second perspective, however, the task of locating drift can be viewed as bearing similarities to the process of enhancing test reliability, since the concept of measurement stability is reflected in both practices in a straightforward manner. When strengthening the reliability of a test, researchers are facing the question of stability of the outcomes – typically the test takers' scores – over repeated administrations which represent the same time point (either the administration of parallel test forms to the same test takers, or a theoretical re-administration of the same test form to the same individuals). For locating

drift, the issue in question is typically the stability of item and test characteristics over multiple administrations which occur at different points in time. According to this view, the drift question lacks the complexity embedded within the concept of validity, and bears similarities to the more simplified practice of evaluating reliability.

Despite its conceptual similarities to reliability and validity, the notion of drift can be viewed as standing on its own under the general umbrella of test stability. With reliability reflecting the quality of the data, or inner-test stability, and validity representing the quality of inferences, or external stability regarding the links between a test and the purpose for which it is used, a lack of drift can be considered the quality of invariance, generalization, and stability of the measure to the passage of time. Sufficient reliability is a precondition to strong validity, since the mere investigation of links between a test and the purpose for its use is meaningless if the test outcomes lack inner-stability. In the same line of thought, strong reliability and validity are preconditions to the aspiration for lack of drift, since the invariance of test characteristics over time is not feasible when the test lacks either inner stability or strong evidence of validity regarding a single time point.

To extend this view of test stability, an additional dimension may be considered: lack of *differential functioning*, or stability of the test characteristics across population groups. This concept can be thought of as a fourth dimension, since the stability of a test in terms of its reliability, validity, and lack of drift, may be different for various groups (e.g., based on demographic variables such as gender, ethnicity, or socioeconomic status). As with the case of drift, the concept of differential functioning is frequently narrowed down to methods which are targeted at evaluating *Differential Item Functioning (DIF)*. The tasks of locating

DIF between two population groups and IPD between two time points are often considered formally identical, since in both cases the issue in question is the comparability of item characteristics across the two groups; and, in both cases, the distributions of test takers' scores regarding the construct being measured cannot be assumed equal across groups. In light of these similarities, practices for locating DIF and IPD share some common methods, with practical differences generally confined to the interpretation of the findings. The view of both concepts as expressing dimensions of test stability supports this similarity.

However, careful examination of the concepts of differential functioning and drift reveals two different measurement narratives. While differential functioning essentially deals with *separating* test data according to the test takers' affiliation with social groups, and evaluating whether the characteristic of interest is held similar across the selected subgroups, the evaluation of drift involves *expanding* the data to other groups of individuals – test takers at different time points – and examining whether a characteristic for one group can be generalized to subsequent groups. This conceptual difference between *inner-projection* and *generalization* of test data entails caution when the same methods are used for detecting both types of lack of invariance regarding test characteristics. This distinction may be clarified following the upcoming part of this conceptualization, in which the notion of drift is decomposed into its basic elements.

Terminology

Throughout this paper, the term “tasks” is used in order to represent the building blocks composing a measurement instrument (often referred to as “items”). This term was

selected due to its generality and verbal meaning, which reminds the readers about its function – an assignment to be completed by those tested by the instrument, which is designed to initiate a response process based on the construct of interest, and may appear in various formats (e.g., multiple choice questions, open-ended questions). The individuals responding to the tasks are referred to as "test takers", a term which signifies subjects possessing the underlying construct, which are being measured by the test of interest.

Measurement over Time

The possibility of drift is feasible when a measurement instrument is used repeatedly over time. More specifically, drift is marked by a change in the characteristics of a measurement instrument; those characteristics derive from the interaction between the instrument's tasks and the test takers who responded to the tasks. Therefore, the three basic elements in the definition of drift are: *tasks*, *test takers*, and *time*. At each time point, the interaction between tasks and test takers produces responses, which are the basis for evaluating the instrument's characteristics of interest. Differences in the characteristics over time may be defined as drift, depending on their nature (for example, size and direction) and the context in which the instrument is being used.

The three basic elements – tasks, test takers, and time – require further specification. Drift can be evaluated across two or more time points, regarding either one specific task, or a measurement instrument composed of any number of tasks. There are two general conditions regarding changes of test takers and tasks over time: the test takers may remain the same or differ, and the tasks may be fixed or changing; and, when different tasks are administered

over time, they can be regarded as comparable (for example, alternate forms), or such comparability may not be assumed. The six situations deriving from all possible combinations of those conditions are summarized in Table 1. Drift is generally associated with situations in which the test takers are different over time (Table 1, condition b). In order to obtain a better understanding of the nature of drift, a discussion regarding all measurement situations mentioned in Table 1 is in place.

Under condition 1a (i.e., the combination of conditions 1 and a), both the test takers and tasks remain the same over time – a measurement framework commonly referred to as *response shift*. This expression reflects the fact that over time, test takers may undergo

Table 1

Frameworks of measurement over time

Tasks	**Test Takers**	
	Condition a	**Condition b**
	Same test takers tested over time	Different test takers tested over time
Condition 1		
Same tasks administered over time	Response shift	Drift (task level)
Condition 2		
Comparable sets of tasks administered over time	Test-retest (alternate forms)	Drift (assessment level)
Condition 3		
Non-comparable sets of tasks administered over time	Longitudinal measurement	-

changes regarding the construct being measured, changes due to the process of repeated measurements, or changes due to exposure to the same specific task repeatedly; tasks, however, remain the same regardless of the time and extent of their administration. Therefore, two non-separable changes are being measured under this condition: the aforementioned changes within the test takers, and changes in the interaction between the test takers and tasks. The definition of this interaction according to the writer's view is elaborated ahead.

Conditions 2a and 3a involve administering different tasks to the same test takers over time. The focus of measurement within such frameworks can be interpreted in several ways depending on the context and assumptions made about task comparability. Beyond measuring the same construct, comparability of tasks or sets of tasks – denoted here by *forms* – can be featured in the following ways. Tasks and/or forms may be assumed comparable either due to their content or to various psychometric properties. In Classical Test Theory, forms are considered psychometrically parallel if they have the same expected true scores and the same observed score variance. Forms are defined as *tau equivalent* if the true scores are assumed equal, and *essentially tau equivalent* if the true score variance is equal across forms. Another type of form comparability – *random equivalence* – can be assumed if the process of form construction involved random selection from a single universe of tasks measuring the same construct. All the cases of form comparability detailed above are represented under condition 2 in Table 1. In any of those cases, the administration of different forms to the same test takers (condition 2a) is generally performed in order to obtain additional measurements of the same construct within the same individuals without presenting the same form twice – or,

in other words, *test-retest* using *alternate forms*. In this framework, the obtained scores are presumed comparable due to the comparability of the forms, and the proximity of the scores can be attributed to reliability (i.e. similar scores indicate high reliability, and vice versa). However, careful examination reveals that differences between the obtained scores may also be influenced by the passage of time, due to some of the factors mentioned for condition 1a: changes within the test takers (regarding the construct being measured, or due to the process of repeated assessments), and changes in the interaction between the test takers and tasks measuring the construct.

Condition 3a, under which different tasks which are designed to measure the same construct are administered over time to the same test takers without the assumption of comparable forms, is known as *longitudinal measurement*. In this case, multiple measurements of the same individuals regarding a single construct are obtained over time, in order to collect information about processes undergone by individuals regarding the construct of interest. Therefore, by definition, the focus in such cases is on changes within the test takers regarding the construct being measured; however, as with the previous conditions, other factors associated with the passage of time also come into play, namely changes due to repeated measurements and changes in the interaction between the test takers and tasks measuring the construct of interest.

Condition 1b specifies a situation in which the same tasks are administered to different test takers over time. This is the typical scenario under which drift is investigated; in such cases, since the tasks do not change by definition, changes in the characteristics of tasks or test forms can be attributed either to the groups of test takers or to the interaction between

the test takers and tasks. Since different individuals are tested over time, the changes between groups of test takers refer to group-level differences with respect to the construct being measured, e.g. due to changes in the population of test takers. In the assessment of drift, the goal is to separate the differences in the interaction between test takers and tasks from differences in the populations of test takers, i.e. to evaluate the changes in the interaction over and above the population differences over time. In other words, the assessment of drift is concerned with answering the following question: do individuals from different time points who are similar with respect to the underlying construct interact differently with the tasks? This choice of wording is intended to reflect the writer's view behind this mysterious "interaction" component: here, "interaction" refers not to the statistical concept regarding two or more variables operating together, but rather to the verbal meaning of a person interacting with a task. In DIF analyses, a finding of such "interaction" component, i.e. differences in performance of individuals from different groups over and above the underlying construct, suggests that several assumptions of the measurement model may have been violated (for example: dimensionality, randomness of errors). In the context of drift, the definition of this interaction as violation of model assumptions is problematic since it implies that when taken together, all test takers from different time points compose a unified population. The writer suggests that the appropriateness of this view in the case of different points in time is questionable, and that the drift scenario is better featured by "subsequent populations of test takers" than by a single population divided into subgroups. Under this formalization, the finding of an interaction component, i.e. drift, can be explained as "differences in the way individuals from different time points react to the task". The scope of

drifting items which measure a common construct can be considered in order to evaluate whether the differences are restricted to the content of few individual items, or indicative of a general change in the definition and manifestation of the construct over time.

Conditions 2b and 3b specify a situation in which different tasks measuring the same construct are administered to different test takers over time. If the sets of tasks – or forms – are assumed to be comparable (condition 2b), drift can be evaluated at the assessment level. That is, using the definitions specified above, when differences in the interaction between test takers and forms are found beyond changes in the populations of test takers, those differences cannot be attributed to specific tasks but to the assessment as a whole, indicating changes in the association between the underlying construct and the assessment's framework. This definition implies a link connecting assessment-level drift and changes in validity over time. Moreover, findings of assessment-level drift for various measures which are designed to capture the same construct may indicate that the changes should be traced back to the construct's definition. When comparability of the forms administered over time is not assumed (condition 3b), the evaluation of drift becomes highly limited due to lack of basis for comparisons.

Readers may note that Table 1 includes three conditions for tasks but only two conditions for test takers. The “missing” condition for test takers is one in which different groups of test takers, which are assumed to be comparable, are tested over time. This condition is not included in the table for two reasons. First, it is the writer's opinion that comparability of groups of test takers regarding a specific construct at different time points cannot be tested. Such examination would require administering either the same test or

measure, or parallel tests which are designed to capture the construct, to both groups; since the two groups will be tested at different time points by definition, this framework will be identical to that of evaluating drift (conditions 1b and 2b). Therefore, it is the writer's logical conclusion that if the possibility of drift is reasonable regarding specific measures and constructs, then comparability of groups of test takers at different time points can only be *assumed*, or at best *supported* using external evidence. The second reason for the absence of an option of test takers comparability in Table 1 refers to the practical definitions of what can be tested under this theoretical scenario. If the groups of test takers are assumed comparable, condition 1b still represents drift at the task level, and condition 2b still stands for drift at the assessment level; the only difference lies in the confidence at which differences in task characteristics can be attributed to drift rather than changes in the population of test takers.

In contrast to comparability of groups of test takers over time, the comparability of tasks or test forms *can* be assessed (using, for example, random assignment of test takers to forms). However, it should be noted that this comparability can only be evaluated regarding a specific point in time; i.e., the fact that two sets of tasks were found comparable at a given point in time does not guarantee that this comparability would hold at subsequent time points. Therefore, the application of condition 2 in Table 1 entails the assumption that form comparability is maintained over time. In the context of drift analyses, this assumption is highly problematic, since it involves the premise of the forms undergoing the same nature of drift.

To summarize, according to the writer's view, drift is defined as changes in the interaction between test takers and measurement tools over time, over and above differences

between groups of test takers at different time points or individual-level changes within test takers. In other words, findings of drift indicate that over time, the definition and manifestation of the construct being measured may be changing. Assessment-level drift can be examined using alternate forms; however, this framework entails the assumption that form comparability is maintained over time. Therefore, for the purpose of tracking drift, condition 1 (the same tasks over time) is less troublesome than condition 2 (comparable tasks; see Table 1). In both measurement frameworks derived from combining condition 1 with condition a or b (same or different test takers over time, respectively), changes in task characteristics over time include drift along with differences either within or between test takers. While comparability of groups of test takers at different time points cannot be proved, it can at least be supported; in contrast, changes which individual test takers may go through over time, and the effect of re-answering the same tasks, seem harder to evaluate. In addition, re-testing the same test takers is often not feasible. Therefore, condition 1b – task-level drift, different test takers over time – seems to be the most appropriate for practical assessment of drift.

Formulation of Drift

Following the conceptualization detailed above, a more specific definition, which may be utilized for the development and demonstration of quantitative methods for assessing drift, is required. As mentioned earlier, drift can be evaluated regarding one or more tasks, and across two or more time points. For task-level drift (condition 1b in Table 1), the basic data matrix consists of task responses denoted by U_{ijk} , where i (= 1, … , m) represents the

task, j (= 1, … , r) indicates the time point, and for each time point j, k (= 1, … , n_j) represents a test taker from the group tested at time j. A visual representation of this matrix is shown in Table 2.

Table 2

Basic data structure for analyses of task-level drift

Tasks	**Time 1**				**Time 2**					**Time *r***			
	TT 1	**TT 2**	. .	**TT n_1**	**TT 1**	**TT 2**	. . . .	**TT n_2**	. . .	**TT 1**	**TT 2**	. .	**TT n_r**
Task 1	$U_{1,1,1}$	$U_{1,1,2}$	. .	$U_{1,1,n1}$	$U_{1,2,1}$	$U_{1,2,2}$	. . . .	$U_{1,2,n2}$	. . .	$U_{1,r,1}$	$U_{1,r,2}$	. .	$U_{1,r,nr}$
Task 2	$U_{2,1,1}$	$U_{2,1,2}$	. .	$U_{2,1,n1}$	$U_{2,2,1}$	$U_{2,2,2}$	. . . .	$U_{2,2,n2}$	. . .	$U_{2,r,1}$	$U_{2,r,2}$	. .	$U_{2,r,nr}$
.	.	.	. .	.	.	.	. . . .	.	. . .	.	.	. .	.
.	.	.	. .	.	.	.	. . . .	.	. . .	.	.	. .	.
.	.	.	. .	.	.	.	. . . .	.	. . .	.	.	. .	.
Task *m*	$U_{m,1,1}$	$U_{m,1,2}$	. .	$U_{m,1,n1}$	$U_{m,2,1}$	$U_{m,2,2}$	. . . .	$U_{m,2,n2}$	. . .	$U_{m,r,1}$	$U_{m,r,2}$	. .	$U_{m,r,nr}$

Note: TT = Test Taker
$U_{i,j,k}$ = Response to task i at time j, by test taker k.

The nature of the data is expressed in the layout of Table 2 in the following way: each row contains all the responses to a single task, and each column includes all the responses given by a single test taker. The number of rows equals the number of tasks (m), and the number of columns corresponds to the total number of test takers. Since it is assumed that different test takers are tested at different time points, the number of columns equals the sum of the number of test takers at each time point:

$$N = \sum_{j=1}^{r} n_j$$

For each time point and task, the characteristic of interest can be thought of as a function of the task responses at the particular time point. More specifically, a characteristic of task i at time j can be defined as

$$c_{ij} = F(\bar{u}_{ij}) \quad ,$$

where

$$\bar{u}_{ij} = (u_{ij1}, u_{ij2}, ..., u_{ijn_j})$$

is a vector of the n_j responses to task i given by the test takers tested at time j. Under this notation, a lack of drift for task i across all r time points is defined as

$$F(\bar{u}_{ij}|j=1) = F(\bar{u}_{ij}|j=2) = ... = F(\bar{u}_{ij}|j=r) \quad ,$$

or

$$c_{i1} = c_{i2} = ... = c_{ir} \quad .$$

A graphical representation of the characteristics derived from task responses at the various time points is shown in Table 3. As in the notation above, the characteristics are denoted in the table by c_{ij} , where i (= 1, … , m) and j (= 1, … , r) indicate the task and time point, respectively. Using this notation, for each task i, c_{i1} , c_{i2} , … , c_{ir} – the task

Table 3

Data structure for analyses of task-level drift – origins of task characteristics

Tasks	**Time 1**				**Time 2**				. . .	**Time *r***			
	TT 1	**TT 2**	. .	**TT n_1**	**TT 1**	**TT 2**	. . .	**TT n_2**	. . .	**TT 1**	**TT 2**	. .	**TT n_r**
Task 1	C_{11}				C_{12}				. . .	C_{1r}			
Task 2	C_{21}				C_{22}				. . .	C_{2r}			
.	.				.				. . .	.			
.	.				.				. . .	.			
.	.				.				. . .	.			
Task *m*	C_{m1}				C_{m2}				. . .	C_{mr}			

Note: TT = Test Taker
C_{ij} = Characteristic of task i at time j.

characteristics at the different time points – are compared; differences are evaluated for significance (statistical and clinical), size and direction, and patterns of differences over time are considered.

For drift at the assessment level (condition 2b in Table 1), the basic data matrix consists of task responses denoted by U_{jik}, where j (= 1, … , r) represents the time point; for each time point j, i (= 1, … , m_j) indicates a task included in the form administered at time j, and k (= 1, … , n_j) represents a test taker from the group tested at that time point. A visual representation of this matrix is provided in Table 4. The nature of the data is expressed in the layout of this table in the following way: each row contains all the responses to a single task, and each column includes all the responses given by a single test taker. As all test takers responded only to the tasks of a unique test form administered at their time point, the table appears to consist of r distinct matrices, where r is the number of time points; or, in other words, much of the potential data is missing by design. The number of rows equals the total number of tasks, which is the sum of the number of tasks at each time point:

$$M = \sum_{j=1}^{r} m_j \quad ;$$

The number of columns corresponds to the total number of test takers, which equals the sum of the number of test takers at each time point:

$$N = \sum_{j=1}^{r} n_j \quad .$$

Table 4

Basic data structure for analyses of assessment-level drift

Test Forms	**Time 1**				**Time 2**						. . .	**Time *r***			
Form 1	TT 1	TT 2	.	TT n_1	TT 1	TT 2	.	.	.	TT n_2	. . .	TT 1	TT 2	.	TT n_r
Task 1	$U_{1,1,1}$	$U_{1,1,2}$	.	$U_{1,1,n1}$											
Task 2	$U_{1,2,1}$	$U_{1,2,2}$	.	$U_{1,2,n1}$											
.	.	.	.	.											
.	.	.	.	.											
.	.	.	.	.											
Task m_1	$U_{1,m1,1}$	$U_{1,m1,2}$	.	$U_{1,m1,n1}$											
Form 2															
Task 1					$U_{2,1,1}$	$U_{2,1,2}$	.	.	.	$U_{2,1,n2}$					
Task 2					$U_{2,2,1}$	$U_{2,2,2}$	.	.	.	$U_{2,2,n2}$					
.					.	.	.	.	.	.					
.					.	.	.	.	.	.					
Task m_2					$U_{2,m2,1}$	$U_{2,m2,2}$	.	.	.	$U_{2,m2,n2}$					
.											.				
.											.				
.											.				
Form *r*															
Task 1												$U_{r,1,1}$	$U_{r,1,2}$	.	$U_{r,1,nr}$
Task 2												$U_{r,2,1}$	$U_{r,2,2}$	.	$U_{r,2,nr}$
.												.	.	.	.
.												.	.	.	.
.												.	.	.	.
.												.	.	.	.
Task m_r												$U_{r,mr,1}$	$U_{r,mr,2}$	.	$U_{r,mr,nr}$

Note: TT = Test Taker
$U_{j,i,k}$ = Response to task *i* from the unique test form administered at time *j*, by test taker *k*.

For each time point, the characteristic of interest can be thought of as a function of the responses to the tasks within the administered form. More specifically, the assessment's characteristic at time *j* can be defined as

$$c_j = F(U_j) \quad ,$$

where

$$U_j = \begin{matrix} u_{j11} & u_{j12} & \dots & u_{j1n_j} \\ u_{j21} & u_{j22} & \dots & u_{j2n_j} \\ \dots & \dots & \dots & \dots \\ u_{jm_j1} & u_{jm_j2} & \dots & u_{jm_jn_j} \end{matrix}$$

is an $m_j \times n_j$ matrix of task responses, with each row containing responses to a single task. Under this notation, a lack of assessment-level drift on all *r* time points is defined as

$$F(U_j|j=1) = F(U_j|j=2) = ... = F(U_j|j=r) \quad ,$$

or

$$c_1 = c_2 = ... = c_r \quad .$$

A graphical representation of the characteristics derived from the responses at the various time points is shown in Table 5. The characteristics are denoted in the table by C_j, where *j* (= 1, … , *r*) indicates the time point. Using this notation, C_1, C_2, … , C_r – the assessment's characteristics at the different time points – are compared; differences are evaluated for significance (statistical and clinical), size, and direction, and patterns of differences over time are considered.

The comparison between form characteristics C_1, C_2, … , C_r (Table 5) for assessment-level drift, and that of task characteristics C_{i1}, C_{i2}, … , C_{ir} for each task *i* in cases of task-level drift (Table 3), is not straightforward. As discussed earlier, changes in the values of a characteristic of interest over time may occur partially due to differences between groups of test takers at the various time points. When group comparability regarding the construct being measured is not assumed or supported, those differences should be taken into account. Methods which allow for comparison while neutralizing group differences have been the focus of psychometric research in the area of drift. These procedures are discussed ahead.

Table 5

Data structure for analyses of assessment-level drift – origin of the assessment's characteristics

Test Forms	**Time 1**	**Time 2**	. . .	**Time *r***
Form 1	TT 1 TT 2 . . TT n_1	TT 1 TT 2 . . . . TT n_2	. . .	TT 1 TT 2 . . TT n_r
Task 1				
Task 2				
.	C_1			
.				
.				
Task m_1				
Form 2				
Task 1				
Task 2				
.				
.		C_2		
Task m_2				
.			.	
.			.	
.			.	
Form *r*				
Task 1				
Task 2				
.				
.				C_r
.				
.				
Task m_r				

Note: TT = Test Taker
C_j = The assessment's characteristic at time j

Methods for Detecting Drift

A review of psychometric literature which deals with the detection of drift reveals several interesting findings. First, the procedures can be divided into two groups: one in which methods for detecting differential functioning are utilized for the evaluation of drift (frequently between two time points), and another which includes methods specifically

designed to assess drift. Second, in contrast to the formal definition and approaches presented earlier in this paper about the nature of drift, the methods generally do not involve a stand-alone description or calculation of the characteristic of interest at various time points; the existing procedures seem to convey a pragmatic approach, in which the question of drift is either narrowed down to whether or not any differences exist, or specific changes over time are being tested (for example, differences which follow a pre-defined functional form). More specifically, the neutralization of group differences is frequently established through the use of a *matching variable*: a variable representing the construct being measured, which is presumed independent of the task under study. The test takers are divided into subgroups according to their position along the matching variable's scale, differences regarding the characteristic of interest are evaluated separately for each subgroup, and the subgroup differences are summarized to an overall difference – to which significance and effect size can be evaluated using statistical tests and procedures. This process entails the assumption that the matching variable has a common scale for all groups – i.e., that identical scores have identical meanings regardless of group membership. In the context of drift analyses, this assumption can be criticized based on the rationale presented earlier in this conceptualization regarding the comparability of groups of test takers over time: placing groups which were tested at different time points on the same scale ignores the possibility of drift in the instrument capturing the matching variable.

Another issue which is noteworthy relates to the nature of what was so far referred to as a 'characteristic of interest'. The methods are mostly designed for investigation of changes in task *difficulty*, and task *discrimination* is sometimes also considered. That is, the

characteristic is narrowed down to one or two specific task qualities. It should be noted, however, that a characteristic of interest – as implied from its choice of wording – is naturally selected based on its practical importance. The use of measurement models in which items are characterized based solely on their difficulty, or on their difficulty and discrimination, is not uncommon; therefore, since the characteristics of interest naturally correspond to the applied measurement model, it is not surprising that task difficulty and discrimination are the most frequently addressed characteristics in the context of drift.

Finally, the conceptual connection between IPD and DIF is expressed in the practice of identifying drift in anchor items during the process of test equating. In this context, changes in task characteristics are often identified across two time points: current and previous administration; the sequential aspect of time is expressed in the larger pattern of the equating design, which is generally aimed at diminishing threats of assessment-level drift across multiple time points. In their discussion about various ways in which drift can be incorporated into the equating process, Arce-Ferrer & Bulut (2016) define IPD as the "phenomenon of change in anchor item parameters across occasions" (p. 2); this definition reflects a pragmatic perception, in which the interest in drift originates in the need to eliminate its effect on equating procedures.

Tests of DIF

In the following part, methods for detecting item-level differential functioning, which may be applicable for evaluating drift under the caveats mentioned earlier, are presented. DIF is mostly defined in the psychometric literature either as differences between statistic/

psychometric properties of an item for groups that are matched on the attribute measured by the test (e.g. Cuevas & Cervantes, 2012; Hidalgo & Lopez-Pina, 2004), or as differences in the probability of responding in a particular category for different groups who are matched on the construct being measured (e.g. Woods, Cai, & Wang, 2013). The second definition, which refers directly to a computational process used for identifying DIF, demonstrates that technique-driven approaches prevail not only in cases of drift analyses but also in the more commonly-addressed realm of differential functioning. The methods for detecting DIF most commonly reported in the psychometric literature are the Mantel-Haenszel (MH) procedure, Logistic Regression (LogR), and various Item Response Theory (IRT)-based methods.

The MH procedure, first introduced by Mantel & Haenszel (1959) in the context of medical research and later developed by Holland & Thayer (1985; 1986), tests whether the odds of a correct response to an item are identical for test takers from the two groups of interest (commonly referred to as the focal and reference groups) along various segments of the matching variable's scale. The procedure includes a Chi-square test of statistical significance and an effect size measure (Zwick, 2012). The MH method is suitable for binary items (scored correct/incorrect), and tests only for uniform DIF, i.e. differences in odds which are similar in size and direction for all segments of the matching variable. It has been criticized as sensitive to the specifically applied division of the matching variable's scale into segments.

As a method for detecting DIF in binary items, LogR involves using regression analysis to predict the logit of correct response to a studied item using the following three nested models. The first model has only the matching variable as a predictor, the second has

both the grouping variable and the matching variable, and the third model has three predictors: the matching variable, the grouping variable, and the interaction between the two. Uniform and non-uniform DIF are identified either on the basis of the regression coefficients obtained in the third model (Swaminathan & Rogers, 1990, p. 363) or with regard to Chi-square tests of model fit for the three models (Zumbo, 1999); the DIF effect size can be evaluated either using the size and direction of the regression coefficients (transformed to an odds metric) or by the differences in R squared between nested models. Zumbo (1999) suggested an extension of the LogR procedure for detecting DIF to ordinal LogR in cases of items that are scored on a likert-type scale, which views the responses as a projection of an underlying continuous variable. The extension builds upon the LogR procedure by replacing each of the three regression models with $\boldsymbol{u} - 1$ models, where $\boldsymbol{u}$ is the number of response categories: in each model, the dependent variable is replaced by the logit of achieving a score (or level of item endorsement) which is at category $\boldsymbol{i}$ or higher ($\boldsymbol{i} = 2, \ldots, \boldsymbol{u}$).

IRT-based methods for assessing DIF involve estimating parameters of the studied item for each group separately, and contrasting either the two sets of parameters or the Item Characteristic Curves (ICCs) deriving from those parameters. Uniform DIF is characterized by ICCs which are shifted horizontally (i.e. differ only with regard to the difficulty parameter), and non-uniform DIF is identified by ICCs which cross one another (Zumbo, 2007, p. 226). Lord (1980, p. 223) proposed a Chi-square statistic for comparing item parameters. More recently, Woods et al. (2013) proposed and tested several approaches to the detection of DIF based on an improved version of Lord's (1980) Chi-square statistic. Raju (1988) developed formulas for calculating either signed or unsigned difference between two

ICCs along the ability continuum, and later (1990) included z-statistics to test for the significance of the differences. The formulas are given under two conditions: difference only between difficulty parameters, and difference between the discrimination and difficulty parameters. Under IRT methodology, these conditions represent uniform and non-uniform DIF, respectively (Zumbo, 2007, p. 226). Kim & Cohen (1991) suggested evaluating the signed or unsigned difference between ICCs only within a trimmed interval of interest on the ability continuum. For models which include a guessing parameter, contrasting ICCs along the entire ability scale is not applicable when the lower asymptotes vary between groups, since the difference between ICCs becomes infinite. More recently, in a discussion about procedures for identifying IPD, Wells et al. (2014) incorporated an additional statistic for summarizing the discrepancy between two ICCs, which is calculated as a weighted sum of the squared differences between two item response functions along a finite number of quadrature points. The proposed index is essentially a test of differences between two groups, which can be used in a wider context of analyzing changes in item characteristics over time. It should be noted that in IRT models, test taker ability is treated as a latent variable which has an arbitrary metric; therefore, comparisons of either estimated parameters or ICCs for different population groups are basically unmatched on ability.

Tests of IPD

With groups of test takers representing time periods, the groups have a natural order, and the interest lies not only in evaluating differences between consecutive time points, but also in finding patterns of differences over time (DeMars, 2004). Documented methods and

procedures tailored to evaluating IPD are detailed ahead.

Bock, Muraki, & Pfeiffenberger (1988) modeled IPD under the 3-Parameter Logistic (3PL) IRT model as either linear or quadratic change in the difficulty parameter. According to the authors, IPD can be confined to item difficulty, since changes in discrimination would be preceded by changes in difficulty level, and changes in lower asymptotes can be assumed small if the test instructions remain the same (p. 277). In the suggested model for linear drift, the difficulty parameter is replaced by a polynomial in a time-point variable: the difficulty parameter remains as the constant term, and the time-point variable denoting the changes in difficulty – equal across all items – has a unique coefficient for each item. The coefficients are common across all time points, and their sum is constrained to equal zero. Therefore, by definition, the model is targeted at changes which are not all in the same direction. In the model of quadratic change in difficulty, a component of the squared time-point variable is added to the linear polynomial, with additional multiplier which is unique for each item (and common across time points) without additional constrains. The authors demonstrated their model using data of 29 items administered as part of the College Board Physics achievement test on five occasions over a period of ten years. Their findings supported a linear trend in difficulty, with several items undergoing non-systematic differences in difficulty over time. The authors thus warned against item calibration based on a single cohort. Another interesting suggestion made by the authors relates to test-level drift: acknowledging that overall change in item difficulty is inseparable from overall ability change in the population of test takers, they propose the idea that average drift for all items be regarded as change in the population.

Kim et al. (1995) developed a procedure for detecting DIF between multiple groups, which could be adapted for locating trends in item parameters over time. Focusing on differences between estimated parameters of the 2-Parameter Logistic (2PL) IRT model, the authors' proposed method can be regarded as an extension of Lord's (1980) Chi-square test comparing the difficulty and discrimination parameters between two groups. The authors state that the method can be adapted to include a comparison of lower asymptote parameters. The method allows its users to define pairs of compared groups; therefore, in the case of IPD, the sequential aspect of time can be conveyed by, for example, contrasting groups at consecutive time points.

Veerkamp & Glas (2000), who focused on item exposure in computerized adaptive testing using the 1- and 3- Parameter Logistic IRT models, proposed a method for detecting IPD using cumulative sums of a modification of Lord's (1980) Chi-square statistic. Aligning with its purpose, the method is designed to check whether an item has become easier and less discriminating over time. The procedure uses the initial estimated parameter(s) as a baseline against which all other parameters at subsequent time points are compared. More specifically, at each time point, the difference between estimated parameters divided by an approximated standard error, is calculated and added to the previously cumulated statistics if it is found larger than zero. For the 3PL IRT model, the standard differences in the difficulty and discrimination parameters are combined to a single statistic which is cumulated across time points. The authors regard the lower asymptote as a constant in the context of adaptive testing, claiming that “guessing … may occur less frequently … because … the items are tailored to the proficiency level of the respondents” (Veerkamp & Glas, 2000, p. 378).

Summary

In this conceptualization, the notion of drift is described from several theoretical perspectives, and a brief summary of documented procedures for the detection of drift is provided. The existing techniques seem to be focused on drift at the item/task rather than test/assessment level, using different groups of test takers (rather than the same individuals) over time – a scenario corresponding to condition 1b in Table 1. The characteristics of interest – mostly task difficulty and discrimination – seem to derive from the applied measurement model, and the methods mostly check for the existence of specific patterns of IPD rather than establishing a stand-alone description of the characteristics' trends over time.

The necessity for obtaining characteristic values which are comparable across time points generally requires scaling or equating procedures. It should be noted that methods of this type frequently involve the use of anchor items which are common across two or more administrations. The methods commonly rely upon the assumption that the anchor items are free of drift; the justification of this assumption requires an assessment of drift among the anchor items, which in turn entails using a set of common items. This circular feature embedded within the assessment of drift supports its perception as a central concept of test stability (along with reliability and validity) given earlier in this conceptualization. Lack of drift can be examined routinely and supported by accumulated evidence, but an ultimate proof of such feature is highly difficult, if not impossible, to obtain.

Proposed Procedure for Evaluating Drift

Introduction

In the following part, a procedure for detecting drift is presented. The method is mostly appropriate for evaluating drift under IRT models, since it involves the assumption that task responses can be estimated based on item and person parameters – an approximation which, in IRT methodologies, can be carried out using an item characteristic function. As in common IRT terminology, the term “task responses” refers here to the responses' scoring as derived from the applied measurement model rather than the responses themselves; for example, in the case of multiple choice questions with five response options and two scoring categories, the “task responses” would be the two categories (namely “correct” and ”incorrect”) rather than the five response options. Therefore, the procedure can be used with multiple types of question formats.

The procedure is adequate for a set of tasks which are administered to different test takers over time. It generally applies to measurement situations corresponding to condition 1b in table 1, although careful examination reveals that the procedure does not strictly involve the assumption of the test takers being grouped together in specific time points. However, cases of non-continuous administration mode would allow for better description of the specific pattern of identified drift.

Steps of the Proposed Procedure

Measurement Data. The data are assumed to follow the structure presented earlier in this paper for task-level drift (Table 2): an $m \times N$ matrix of task responses, where m = the number of tasks, and N = the total number of test takers at all time points.

Procedure's Steps. The steps of the proposed procedure are presented in Table 6. The suggested process includes four stages: estimation of model parameters, simulation of reference responses, evaluation of the similarity between the original and reference responses followed by a conclusion regarding drift for each task, and further analyses of tasks with indication of drift. An elaborate discussion of the steps is provided ahead.

Table 6

Steps of the proposed procedure for evaluating drift

Steps	Description	Comments
Step 1.	Use task responses from all time points stacked together in order to estimate model parameters for all tasks and test takers.	The number and nature of the estimated parameters are expected to vary according to the applied measurement model. For example, in case of the unidimensional 3PL IRT model, step 1 would include estimating a total of $N + 3m$ parameters: one parameter for each of the N test takers (latent ability; commonly denoted by θ), and three parameters for each of the m tasks (discrimination, difficulty, and lower asymptote; commonly denoted by a, b, and c, respectively).
Step 2.	Simulate responses to all tasks by all test takers based on the parameters estimated in step 1.	For each task, the simulated responses – referred to as ***Reference Responses*** – represent an alternative response pattern without IPD.
Step 3.	For each task, compare the original and reference responses, and evaluate whether the discrepancy is large enough to conclude that the task exhibits drift.	
Step 4.	For each task with indication of drift, make further analyses in order to evaluate the nature of changes over time.	

Discussion

The procedure begins by stacking all time points together for the purpose of concurrent calibration of all task and test taker parameters. The procedure's first step thus operates under the assumption that for all tasks, a single set of task parameters lies in the basis of all responses to the task regardless of the time in which the task was administered. In other words, the method tests the hypothesis of exchangeability of test taker responses across time points. This initial process results in estimated parameters which are guaranteed to be on the same scale. If the model fits the data and no drift is present in any of the tasks, then the estimated task parameters should approximate the true non-variant parameters, and group-level differences between estimated test taker parameters at different time points would reflect true differences between test takers over time.

By definition, since a single set of parameters is assumed to be associated with the task responses at all time points, the parameters estimated at Step 1 would be ones under which the task responses are most likely to be observed **given the assumption of no drift**. In order to evaluate the plausibility of this assumption, the procedure's second step involves simulating alternative responses to all tasks by all test takers based on the parameters estimated in Step 1. The idea behind this simulation is that for tasks without drift, the alternative responses should be in close proximity to the original data upon which the estimation was based. From this point of the paper onwards, the alternative responses would be referred to as ***Reference Responses***; this notation is intended to emphasize these responses' functionality as data to refer to for evaluating drift in the original data.

In Step 3, the original and reference responses are compared for each task. If the

comparison reveals a discrepancy between the two sets of responses which is beyond chance, the hypothesis of exchangeability of test taker responses between time points is rejected, and the task under consideration is regarded as exhibiting drift. The indicator for comparison between the two sets of responses, and the cutoff for determining a level of discrepancy which is beyond chance, may vary according to the applied measurement model. In all cases, however, the comparison should be conducted under the conceptual understanding that the two sets of responses are *paired* – that is, each reference response is associated with a specific original response, and the association is strong in the sense that both responses stem from the same individual test taker.

If the comparison results in indication of drift, the procedure continues to its last stage (Step 4), which involves conducting further analyses of tasks with identified drift in order to evaluate the size, direction, and possible pattern of changes over time. Readers should note that the time points in which the original data were collected are overlooked throughout the entire process up until this last step. Although the proposed procedure could be applied in cases of continuous administration, the larger the number of test takers at each time point – the easier the task of locating patterns of drift over time. Readers should also note that the parameters upon which the reference responses are simulated, are estimated based on the assumption of lack of drift; therefore, the larger the number of tasks with identified drift, the larger the threat to the reliability of the entire analysis. Such cases would entail further investigation, perhaps by re-estimating model parameters without the drifting tasks.

Several comments regarding the proposed procedure are in place. First, its main strength seems to be the abstention from comparing parameters directly – a comparison

which may be misleading due to parameters that are not on the same scale. The procedure bypasses this problem, as well as the tendency to draw false conclusions about group differences based on non-comparable parameters, by comparing sets of responses rather than parameters. Second, the procedure is generic in the sense that it checks for measurement invariance regardless of the specific concern for its violation. This all-encompassing feature can be regarded as a weakness, since the sequential component of time embedded in the concept of drift, as well as its aforementioned separation from the concept of DIF, are not represented in the procedure's steps.

Lastly, the procedure in its current state presents a general approach to assessing drift, without specific details on how each step should be conducted in practice. That is, Step 1 and Step 2 do not specify methods for estimating neither the test taker and task parameters nor task responses based on those parameters; Step 3 does not provide details about how to conduct the comparison of the original and reference responses, and no decision rules relating this comparison to an overall evaluation of drift are specified; and finally, Step 4 does not elaborate on how the tasks with identified drift may be further analyzed in order to uncover the nature of changes over time. At present, this vagueness is intended to be viewed as allowing for flexibility in various possible implementations of the proposed procedure.

Method

In the following part, an attempt to provide a proof of concept to the proposed procedure for detecting drift is presented. In this demonstration, IPD is evaluated in simulated large scale assessment tasks (see full data description ahead) using the proposed procedure as well as two existing methods: the multiple-groups DIF method described by Kim et al. (1995), and logistic regression. Information about both methods is provided earlier in this paper, and specifications regarding their current application are given ahead. The rates of detection of true drift, drift which was not detected, and detection of drift when it was not present in the data, are compared across the two known methods and the proposed procedure.

Data

The simulated data used in this demonstration included binary responses (e.g., correct / incorrect) to a hypothetical 30-task test form at three time points, which were generated using the 2PL IRT model. The simulation was conducted using WinGen software (Han, 2007). The simulated data were intended to correspond to condition 1b in Table 1, with the test form administered to a different group of 50,000 test takers at each time point. For the first time point, true difficulty and discrimination parameters (denoted by b and a, respectively) were sampled for each of the 30 tasks using the distributions N(0,1) for difficulty and N(1,0.5) for discrimination, and true test taker ability parameters (commonly denoted by theta, or θ) were created based on the standard normal distribution N(0,1). For the second and third time points, six of the 30 tasks were randomly selected as tasks undergoing

changes in one or both parameters over time: two tasks with the difficulty parameter decreasing by 0.4, two tasks with the discrimination parameter decreasing by 0.2, and two tasks with the difficulty parameter decreasing by 0.4 *and* the discrimination parameters decreasing by 0.2. Each type of change was consistent across the two pairs of consecutive time points (i.e., changes from Time 1 to Time 2 were continued in the same size and direction between Time 2 and Time 3). The responses were generated under two conditions: without changes in the distribution of test taker abilities over time, and with the mean test taker ability increasing by 0.1 in the second and third administration (without changes in standard deviations). The full layout of true parameters for all tasks in each administration is presented in Table 7, and the means and standard deviations of true test taker abilities under each condition (with and without changes in ability distribution over time) are presented in Table 8.

The distributions of initial task parameters and the size of the differences in parameters over time were selected in accordance with Rudner, Getson, & Knight (1980) and Donoghue & Isham (1998), respectively. The directions of changes in task parameters – decrease in difficulty and/or decrease in discrimination – were selected based on the rationale that they represent types of drift which may be a source of concern in real measurement frameworks. Over time, a decrease in task discrimination reflects a weakened connection between the task and the construct being measured; a decrease in difficulty may happen, for example, as a result of exposure of tasks in the context of high stakes tests. The direction of changes in test taker parameters – increase in mean ability – may be observed in the context of tests which, over time, become a focus of increased social interest, or due to the impact of

increased preparation for high stakes tests.

Table 7

Task parameters in each simulated administration

Tasks	Parameter(s) changing over time	Time 1		Time 2		Time 3	
		a	***b***	***a***	***b***	***a***	***b***
Task 1	-	1.44	-1.05	1.44	-1.05	1.44	-1.05
Task 2	-	1.77	0.83	1.77	0.83	1.77	0.83
Task 3	-	1.13	-0.14	1.13	-0.14	1.13	-0.14
Task 4	-	1.12	-0.22	1.12	-0.22	1.12	-0.22
Task 5	***a***	0.96	-1.04	**0.76**	-1.04	**0.56**	-1.04
Task 6	-	0.58	0.01	0.58	0.01	0.58	0.01
Task 7	-	1.07	-0.88	1.07	-0.88	1.07	-0.88
Task 8	-	0.93	-0.24	0.93	-0.24	0.93	-0.24
Task 9	-	1.20	-1.43	1.20	-1.43	1.20	-1.43
Task 10	-	1.92	-0.90	1.92	-0.90	1.92	-0.90
Task 11	-	0.57	-0.90	0.57	-0.90	0.57	-0.90
Task 12	***a*** , ***b***	0.48	0.01	**0.28**	**-0.39**	**0.08**	**-0.79**
Task 13	-	1.38	0.82	1.38	0.82	1.38	0.82
Task 14	-	1.11	-1.20	1.11	-1.20	1.11	-1.20
Task 15	***b***	0.28	0.09	0.28	**-0.31**	0.28	**-0.71**
Task 16	-	1.36	1.22	1.36	1.22	1.36	1.22
Task 17	-	0.25	0.35	0.25	0.35	0.25	0.35
Task 18	-	0.82	-0.18	0.82	-0.18	0.82	-0.18
Task 19	-	1.33	0.04	1.33	0.04	1.33	0.04
Task 20	-	1.27	-0.08	1.27	-0.08	1.27	-0.08
Task 21	-	0.96	1.17	0.96	1.17	0.96	1.17
Task 22	-	0.63	-0.88	0.63	-0.88	0.63	-0.88
Task 23	***a***	1.25	1.85	**1.05**	1.85	**0.85**	1.85
Task 24	-	1.52	-0.44	1.52	-0.44	1.52	-0.44
Task 25	-	1.06	-1.19	1.06	-1.19	1.06	-1.19
Task 26	***b***	0.63	-1.09	0.63	**-1.49**	0.63	**-1.89**
Task 27	-	0.87	0.08	0.87	0.08	0.87	0.08
Task 28	***a*** , ***b***	1.95	-0.97	**1.75**	**-1.37**	**1.55**	**-1.77**
Task 29	-	0.48	0.53	0.48	0.53	0.48	0.53
Task 30	-	2.61	0.94	2.61	0.94	2.61	0.94
	Mean	**1.10**	**-0.16**	**1.07**	**-0.22**	**1.04**	**-0.27**
	SD	**0.53**	**0.85**	**0.53**	**0.88**	**0.54**	**0.93**

Note: *a* = Discrimination parameter
b = Difficulty parameter

Table 8

Distributions of test takers' true abilities in each simulated administration, under two conditions: with and without changes in ability distribution over time

Statistics of test taker ability (θ)	Time 1	No changes in ability distribution		Ability distribution changing	
		Time 2	Time 3	Time 2	Time 3
N	50,000	50,000	50,000	50,000	50,000
Mean	0.002	0.003	-0.002	0.105	0.196
SD	1.003	0.998	0.996	0.996	1.002

Data Analysis Process & Results

Detection of IPD using Multiple-Groups DIF (MGDIF)

The MGDIF method described by Kim et al. (1995) involves using an extension of Lord's (1980) Chi-square statistic to indicate the existence of overall DIF among multiple groups (Chan, Drasgow, & Sawin, 1999). The method as specified by Kim et al. (1995) is based on a comparison of the 2PL IRT difficulty and discrimination parameters obtained for each of the groups. The current analysis utilized the method's implementation in R software package difR (Magis et al., 2010). Although the difR package includes options of adjustment for multiple comparisons, in the current application – with only three groups – the results with and without adjustment were very similar, and therefore only the non-adjusted results are reported ahead.

The results of the MGDIF analysis for each of the two conditions in the simulated data (with and without changes in test taker ability distribution over time) are summarized in Table 9 (see R code in Appendix 1). The results indicate that both with and without changes in the test taker ability distribution over time, drift was identified with significance level of .01 for all tasks except Task 6 and Task 17 (due to the large sample size, only a significance level of .01 – rather than .05 – was considered). All tasks with true changes in parameters over time were detected as drifting, but so were most of the non-drifting tasks. It is possible that the high detection of drift is related to the large sample size (N = 50,000 at each time point), and that the effect sizes of the differences for the non-drifting tasks were quite small. This conjecture is strengthened by careful examination of the specific values of the test

Table 9

Detection of IPD using Multiple-Groups DIF (MGDIF), based on the method presented by Kim et al. (1995): Chi-squares and p-values for each condition of the simulated data

Task	Parameter(s) changing over time	No changes in ability distribution		Ability distribution changing	
		χ^2	*p*	χ^2	*p*
Task 1	-	63.54	<.001	91.62	<.001
Task 2	-	45.96	<.001	47.19	<.001
Task 3	-	39.93	<.001	40.28	<.001
Task 4	-	73.78	<.001	52.45	<.001
Task 5	***a***	1024.60	<.001	1116.84	<.001
Task 6	-	5.43	.246	5.22	.266
Task 7	-	75.28	<.001	42.46	<.001
Task 8	-	56.72	<.001	43.84	<.001
Task 9	-	66.19	<.001	63.57	<.001
Task 10	-	103.48	<.001	112.50	<.001
Task 11	-	25.83	<.001	31.29	<.001
Task 12	***a*** , ***b***	688.37	<.001	868.19	<.001
Task 13	-	16.63	.002	40.66	<.001
Task 14	-	90.18	<.001	61.90	<.001
Task 15	***b***	223.03	<.001	182.60	<.001
Task 16	-	33.81	<.001	24.00	<.001
Task 17	-	6.87	.143	7.11	.130
Task 18	-	38.93	<.001	53.78	<.001
Task 19	-	34.22	<.001	48.22	<.001
Task 20	-	49.57	<.001	34.66	<.001
Task 21	-	18.44	.001	18.10	.001
Task 22	-	33.77	<.001	38.59	<.001
Task 23	***a***	722.61	<.001	733.78	<.001
Task 24	-	76.84	<.001	74.69	<.001
Task 25	-	82.75	<.001	41.19	<.001
Task 26	***b***	784.14	<.001	789.65	<.001
Task 27	-	24.92	<.001	14.52	.006
Task 28	***a*** , ***b***	1371.33	<.001	1155.94	<.001
Task 29	-	21.16	<.001	17.10	.002
Task 30	-	26.48	<.001	29.60	<.001

Note: *a* = Discrimination parameter, *b* = Difficulty parameter
Degrees of freedom for Chi-square tests = 4

statistic. The distributions of Chi-square values for the 30 tasks under both conditions are summarized in Appendix 2 using Boxplots; in this summary, without changes in the test taker ability distribution over time, values starting at 223.03 (Task 15) were identified as outliers,

and values staring at 182.60 (Task 15) appeared as outliers when the test taker ability distribution was changing over time. When only the outlier Chi-square values are considered as indicating drift, the MGDIF method results in perfect identification of all tasks with true drift under both conditions. In other words, when the Chi-square values are analyzed critically without the traditional application of p-values, a conclusion featured by the absence of false positives and high power can be reached.

Detection of IPD using Logistic Regression (LogR)

The evaluation of DIF for a specific task using binary LogR involves the use of three variables: a matching variable (indicator of test taker ability), a grouping variable, and the interaction between the two. An analysis of drift can be performed using a grouping variable which is ordinal, with the number of possible values equal to the number of time points. In the current case, for each test taker and task, a "rest score" – total score excluding the analyzed task – was used as the matching variable. For each condition (with and without changes in the distribution of test taker abilities over time), drift was evaluated for each task separately using the three following regression models for predicting the logit of correct response to the task:

Model 1. Independent variable: Rest Score

Model 2. Independent variables: Rest Score, Time Point

Model 3. Independent variables: Rest Score, Time Point, interaction (Rest Score x Time Point)

Three Chi-square tests of differences between nested models were considered: difference between Model 2 and Model 1 (two degrees of freedom), difference between

Model 3 and Model 2 (two degrees of freedom), and difference between Model 3 and Model 1 (four degrees of freedom). The identification of IPD was conducted using the following guidelines:

- A case of significant differences between all three pairs of models was regarded as indicating non-uniform drift which is featured by changes in both the difficulty and discrimination parameters. **Otherwise, the decision rules below were followed.**
- A case of significant difference only between Model 2 and Model 1 was considered as indicating uniform drift (changes only in the difficulty parameter). If, in addition, the difference between Model 3 and Model 1 was significant, the same conclusion was reached.
- A case of significant difference only between Model 3 and Model 2 was regarded as indicating non-uniform drift with changes only in the discrimination parameter. If, in addition, the difference between Model 3 and Model 1 was significant, the same conclusion was reached.
- A case of significant difference only between Model 3 and Model 1 was considered as an overall indication of drift which cannot be attributed to either the difficulty or discrimination parameter.[1]

Due to the large sample size, a significance level of .01 (rather than .05) was used. The analyses were conducted using SPSS statistical software[2] (see sample of the syntax code of

---

1 The missing case of non-significant difference only between Model 3 and Model 1 is not mentioned, since this situation was not observed in the analysis of the current data. In principle, this situation may represent unique cases of drift which require further investigation.

2 IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp

this analysis in Appendix 3). The results of the LogR drift analysis for each of the two conditions in the simulated data (with and without changes in test taker ability distribution over time) are summarized in Table 10.

The results indicate that under both conditions (with and without changes in test taker ability distribution over time), the LogR method was successful in identifying all cases of true changes in either the difficulty or the discrimination parameter. However, in all cases of true changes in the discrimination parameter alone, drift in the difficulty parameter was also identified, and in one case of true changes only in the difficulty parameter – changes in discrimination were also recognized.

Under both conditions (with and without changes in the test taker ability distribution over time), tasks without true drift were often identified as exhibiting either uniform or non-uniform drift; however, it should be noted that the Chi-square values for those cases were notably lower in comparison to the values found for tasks with true drift. An additional finding which may be of interest refers to the type of falsely identified drift under the two conditions. When the test taker ability distribution was changing over time, all such cases resulted in identification of non-uniform drift with changes only in the discrimination parameter; however, when the test taker ability distribution was not changing over time, almost all falsely identified drift was uniform, i.e., targeted at the difficulty parameter.

As in the MGDIF analysis, it is possible that the large sample size contributed to the high rates of false positives, and a critical examination of the Chi-square values may lead to different conclusions. The distribution of Chi-square values for all three tests of differences between nested models are summarized in Appendix 4 using Boxplots. In this summary, for

Table 10

Detection of IPD using logistic regression, for each condition of the simulated data

a. No changes in test taker ability distribution over time

Task	Parameter(s) changing over time	Drift identified	Test of Uniform Drift: Difference between Model 2 and Model 1 (*df*=2)		Tests of Non-Uniform Drift: Difference between Model 3 and Model 2 (*df*=2)		Tests of Non-Uniform Drift: Difference between Model 3 and Model 1 (*df*=4)	
			χ^2	*p*	χ^2	*p*	χ^2	*p*
Task 1	-	Uniform	18.07	<.001	0.01	.995	18.08	.001
Task 2	-	Overall – ?	8.49	.014	7.59	.022	16.08	.003
Task 3	-	-	8.11	.017	4.59	.101	12.70	.013
Task 4	-	Uniform	17.04	<.001	9.26	.010	26.29	<.001
Task 5	***a***	Non-uniform (a,b)	738.54	<.001	427.04	<.001	1165.57	<.001
Task 6	-	-	0.68	.711	0.42	.811	1.10	.894
Task 7	-	Non-uniform (a,b)	17.39	<.001	10.95	.004	28.34	<.001
Task 8	-	Uniform	19.58	<.001	3.58	.167	23.16	<.001
Task 9	-	Uniform	15.31	<.001	7.10	.029	22.41	<.001
Task 10	-	Uniform	29.20	<.001	5.99	.050	35.19	<.001
Task 11	-	Uniform	12.38	.002	3.24	.198	15.62	.004
Task 12	***a*** , ***b***	Non-uniform (a,b)	51.57	<.001	669.09	<.001	720.67	<.001
Task 13	-	-	2.04	.361	2.24	.325	4.28	.369
Task 14	-	Uniform	25.76	<.001	7.34	.026	33.10	<.001
Task 15	***b***	Uniform	253.08	<.001	3.49	.175	256.57	<.001
Task 16	-	-	7.53	.023	4.39	.111	11.91	.018
Task 17	-	-	4.32	.115	0.23	.893	4.54	.337
Task 18	-	Uniform	12.18	.002	5.32	.070	17.50	.002
Task 19	-	-	5.52	.063	1.23	.541	6.75	.150
Task 20	-	-	8.87	.012	3.53	.171	12.39	.015
Task 21	-	-	3.07	.215	1.21	.546	4.29	.369
Task 22	-	Uniform	10.67	.005	1.84	.399	12.51	.014
Task 23	***a***	Non-uniform (a,b)	698.57	<.001	232.00	<.001	930.57	<.001
Task 24	-	Uniform	21.56	<.001	6.37	.041	27.93	<.001
Task 25	-	Uniform	26.29	<.001	6.96	.031	33.25	<.001
Task 26	***b***	Uniform	978.40	<.001	2.92	.232	981.32	<.001
Task 27	-	-	6.05	.049	1.08	.581	7.13	.129
Task 28	***a*** , ***b***	Non-uniform (a,b)	2771.90	<.001	103.54	<.001	2875.44	<.001
Task 29	-	Uniform	12.24	.002	0.74	.690	12.98	.011
Task 30	-	-	4.61	.100	1.33	.515	5.93	.204

Note: 1. *a* = Discrimination parameter, *b* = Difficulty parameter
2. Significance level for identification of drift = .01
3. Variables used as predictors in the nested models:
   Model 1. Rest Score
   Model 2. Rest Score, Time Point
   Model 3. Rest Score, Time Point, interaction between Rest Score and Time Point

(continued in the next page)

Table 10 (continued)

b. Test taker ability distribution changing over time

Task	Parameter(s) changing over time	Drift identified	Test of Uniform Drift		Tests of Non-Uniform Drift			
			Difference between Model 2 and Model 1 (*df*=2)		Difference between Model 3 and Model 2 (*df*=2)		Difference between Model 3 and Model 1 (*df*=4)	
			χ^2	*p*	χ^2	*p*	χ^2	*p*
Task 1	-	-	7.79	.020	4.12	.127	11.92	.018
Task 2	-	Non-uniform (a)	1.75	.417	11.93	.003	13.68	.008
Task 3	-	-	2.66	.264	4.86	.088	7.53	.111
Task 4	-	Non-uniform (a)	2.68	.262	14.84	.001	17.51	.002
Task 5	***a***	Non-uniform (a,b)	754.70	<.001	436.53	<.001	1191.23	<.001
Task 6	-	-	0.74	.692	0.80	.671	1.53	.821
Task 7	-	Non-uniform (a)	0.37	.831	9.56	.008	9.93	.042
Task 8	-	-	5.63	.060	0.69	.710	6.32	.177
Task 9	-	-	3.36	.186	3.16	.206	6.52	.164
Task 10	-	-	6.28	.043	5.57	.062	11.85	.019
Task 11	-	-	3.54	.170	4.46	.108	8.00	.092
Task 12	***a*** , ***b***	Non-uniform (a,b)	69.88	<.001	722.55	<.001	792.43	<.001
Task 13	-	Non-uniform (a)	4.64	.098	17.95	<.001	22.60	<.001
Task 14	-	-	2.38	.303	9.05	.011	11.43	.022
Task 15	***b***	Uniform	254.78	<.001	1.74	.418	256.53	<.001
Task 16	-	Non-uniform (a)	0.10	.950	11.40	.003	11.50	.021
Task 17	-	-	4.90	.086	0.50	.780	5.40	.249
Task 18	-	Non-uniform (a)	5.28	.071	11.30	.004	16.58	.002
Task 19	-	Non-uniform (a)	2.90	.235	11.98	.003	14.87	.005
Task 20	-	-	0.54	.762	2.87	.238	3.41	.491
Task 21	-	-	2.37	.306	4.19	.123	6.56	.161
Task 22	-	-	8.07	.018	1.42	.493	9.49	.050
Task 23	***a***	Non-uniform (a,b)	806.05	<.001	241.89	<.001	1047.93	<.001
Task 24	-	-	2.19	.334	7.72	.021	9.91	.042
Task 25	-	-	0.35	.838	2.72	.257	3.07	.546
Task 26	***b***	Non-uniform (a,b)	1068.17	<.001	12.85	.002	1081.02	<.001
Task 27	-	-	1.52	.468	0.63	.729	2.15	.708
Task 28	***a*** , ***b***	Non-uniform (a,b)	2723.29	<.001	76.12	<.001	2799.41	<.001
Task 29	-	-	4.05	.132	0.09	.956	4.14	.387
Task 30	-	Non-uniform (a)	4.73	.094	10.18	.006	14.91	.005

Note: 1. *a* = Discrimination parameter, *b* = Difficulty parameter
2. Significance level for identification of drift = .01
3. Variables used as predictors in the nested models:
Model 1. Rest Score
Model 2. Rest Score, Time Point
Model 3. Rest Score, Time Point, interaction between Rest Score and Time Point

the tests of uniform drift (Appendix 4a) with changes in the ability distribution over time, values starting at 69.88 (Task 12) were identified as outliers; for tests of uniform drift without changes in the ability distribution over time, values starting at 253.08 (Task 15) emerged as outliers. For the two degrees of freedom tests of non-uniform drift (Appendix 4b), values

starting at 76.12 (Task 28) were identified as outliers with changes in the ability distribution over time, and values starting at 103.54 (Task 28) were recognized as outliers when the ability distribution was not changing over time. For the four degrees of freedom tests of non-uniform drift (Appendix 4c), values starting at 256.53 and 256.57 (both attributed to Task 15) appeared as outliers with and without changes in the ability distribution over time, respectively.

A summary of the identification of drift driven by the outlier Chi-square values (rather than p-values smaller than .01) is presented in Table 11. The results indicate that under both conditions (with and without changes in the test taker ability distribution over time), the LogR method was successful in identifying all cases of true changes in the discrimination parameter, and almost all true changes in difficulty (all except Task 12, which had true drift in both difficulty and discrimination, but was identified as undergoing changes only in discrimination when the ability distribution was not changing over time). However, in all cases of true changes in the discrimination parameter alone, drift in the difficulty parameter was also identified. There were no cases on non-drifting tasks falsely identified with drift of any kind. To summarize, a critical evaluation of the Chi-square values resulted in perfect identification of the truly-drifting tasks, with several cases of false positives and high power regarding the types of identified drift.

Table 11

Identification of IPD using logistic regression for each condition of the simulated data (detection according to extreme values of Chi-square tests of differences between nested models)

a. No changes in test taker ability distribution over time

Task	Parameter(s) changing over time	Drift identified	Test of Uniform Drift: Difference between Model 2 and Model 1 (*df*=2)		Tests of Non-Uniform Drift: Difference between Model 3 and Model 2 (*df*=2)		Difference between Model 3 and Model 1 (*df*=4)	
			χ^2	*p*	χ^2	*p*	χ^2	*p*
Task 1	-	-	18.07	<.001	0.01	.995	18.08	.001
Task 2	-	-	8.49	.014	7.59	.022	16.08	.003
Task 3	-	-	8.11	.017	4.59	.101	12.70	.013
Task 4	-	-	17.04	<.001	9.26	.010	26.29	<.001
Task 5	***a***	Non-uniform (a,b)	738.54	<.001	427.04	<.001	1165.57	<.001
Task 6	-	-	0.68	.711	0.42	.811	1.10	.894
Task 7	-	-	17.39	<.001	10.95	.004	28.34	<.001
Task 8	-	-	19.58	<.001	3.58	.167	23.16	<.001
Task 9	-	-	15.31	<.001	7.10	.029	22.41	<.001
Task 10	-	-	29.20	<.001	5.99	.050	35.19	<.001
Task 11	-	-	12.38	.002	3.24	.198	15.62	.004
Task 12	***a*** , ***b***	Non-uniform (a)	51.57	<.001	669.09	<.001	720.67	<.001
Task 13	-	-	2.04	.361	2.24	.325	4.28	.369
Task 14	-	-	25.76	<.001	7.34	.026	33.10	<.001
Task 15	***b***	Uniform	253.08	<.001	3.49	.175	256.57	<.001
Task 16	-	-	7.53	.023	4.39	.111	11.91	.018
Task 17	-	-	4.32	.115	0.23	.893	4.54	.337
Task 18	-	-	12.18	.002	5.32	.070	17.50	.002
Task 19	-	-	5.52	.063	1.23	.541	6.75	.150
Task 20	-	-	8.87	.012	3.53	.171	12.39	.015
Task 21	-	-	3.07	.215	1.21	.546	4.29	.369
Task 22	-	-	10.67	.005	1.84	.399	12.51	.014
Task 23	***a***	Non-uniform (a,b)	698.57	<.001	232.00	<.001	930.57	<.001
Task 24	-	-	21.56	<.001	6.37	.041	27.93	<.001
Task 25	-	-	26.29	<.001	6.96	.031	33.25	<.001
Task 26	***b***	Uniform	978.40	<.001	2.92	.232	981.32	<.001
Task 27	-	-	6.05	.049	1.08	.581	7.13	.129
Task 28	***a*** , ***b***	Non-uniform (a,b)	2771.90	<.001	103.54	<.001	2875.44	<.001
Task 29	-	-	12.24	.002	0.74	.690	12.98	.011
Task 30	-	-	4.61	.100	1.33	.515	5.93	.204

Note: 1. *a* = Discrimination parameter, *b* = Difficulty parameter
2. Variables used as predictors in the nested models:
Model 1. Rest Score
Model 2. Rest Score, Time Point,
Model 3. Rest Score, Time Point, interaction between Rest Score and Time Point

(continued in the next page)

Table 11 (continued)

b. Test taker ability distribution changing over time

Task	Parameter(s) changing over time	Drift identified	Test of Uniform Drift: Difference between Model 2 and Model 1 (df=2) χ^2	p	Tests of Non-Uniform Drift: Difference between Model 3 and Model 2 (df=2) χ^2	p	Difference between Model 3 and Model 1 (df=4) χ^2	p
Task 1	-	-	7.79	.020	4.12	.127	11.92	.018
Task 2	-	-	1.75	.417	11.93	.003	13.68	.008
Task 3	-	-	2.66	.264	4.86	.088	7.53	.111
Task 4	-	-	2.68	.262	14.84	.001	17.51	.002
Task 5	***a***	Non-uniform (a,b)	754.70	<.001	436.53	<.001	1191.23	<.001
Task 6	-	-	0.74	.692	0.80	.671	1.53	.821
Task 7	-	-	0.37	.831	9.56	.008	9.93	.042
Task 8	-	-	5.63	.060	0.69	.710	6.32	.177
Task 9	-	-	3.36	.186	3.16	.206	6.52	.164
Task 10	-	-	6.28	.043	5.57	.062	11.85	.019
Task 11	-	-	3.54	.170	4.46	.108	8.00	.092
Task 12	***a*** , ***b***	Non-uniform (a,b)	69.88	<.001	722.55	<.001	792.43	<.001
Task 13	-	-	4.64	.098	17.95	<.001	22.60	<.001
Task 14	-	-	2.38	.303	9.05	.011	11.43	.022
Task 15	***b***	Uniform	254.78	<.001	1.74	.418	256.53	<.001
Task 16	-	-	0.10	.950	11.40	.003	11.50	.021
Task 17	-	-	4.90	.086	0.50	.780	5.40	.249
Task 18	-	-	5.28	.071	11.30	.004	16.58	.002
Task 19	-	-	2.90	.235	11.98	.003	14.87	.005
Task 20	-	-	0.54	.762	2.87	.238	3.41	.491
Task 21	-	-	2.37	.306	4.19	.123	6.56	.161
Task 22	-	-	8.07	.018	1.42	.493	9.49	.050
Task 23	***a***	Non-uniform (a,b)	806.05	<.001	241.89	<.001	1047.93	<.001
Task 24	-	-	2.19	.334	7.72	.021	9.91	.042
Task 25	-	-	0.35	.838	2.72	.257	3.07	.546
Task 26	***b***	Uniform	1068.17	<.001	12.85	.002	1081.02	<.001
Task 27	-	-	1.52	.468	0.63	.729	2.15	.708
Task 28	***a*** , ***b***	Non-uniform (a,b)	2723.29	<.001	76.12	<.001	2799.41	<.001
Task 29	-	-	4.05	.132	0.09	.956	4.14	.387
Task 30	-	-	4.73	.094	10.18	.006	14.91	.005

Note: 1. a = Discrimination parameter, b = Difficulty parameter
2. Variables used as predictors in the nested models:
Model 1. Rest Score
Model 2. Rest Score, Time Point,
Model 3. Rest Score, Time Point, interaction between Rest Score and Time Point

Detection of IPD using the Procedure Proposed in this Paper

The analysis of IPD according to the procedure proposed earlier in this paper was operationalized for each condition of the simulated data (with and without changes in the test taker ability distribution over time) in the following way. First, task parameters and abilities

were estimated for all tasks and test takers according to the 2PL IRT model, based on all responses from the three times points stacked together. This initial step was conducted using jMetrik software (Meyer, 2014). In this procedure, Beta distribution β(1.75, 3) with a minimum of 0 and maximum of 3 was used as a prior for discrimination, and Beta distribution β(1.01, 1.01) ranging between -6 and 6 was applied as a prior for difficulty. Expected a posteriori ability estimation was applied, and the distribution of estimated test taker ability was scaled to a mean of 0, standard deviation of 1, and a range of -6 to 6. The estimated task parameters and statistics of the test taker estimated ability distributions under the two conditions, are available in Appendix 5 and Appendix 6, respectively.

The second step involved sampling an alternative binary response (correct / incorrect) to each task by each test taker based on the 2PL IRT model, using the task and test taker parameters estimated in Step 1. As noted in the description of the proposed procedure earlier in this paper, these reference responses represent a possible response pattern to the tasks given the assumption that all tasks were free of drift. This simulation was conducted using the same software which was applied for the initial data generation (WinGen; Han, 2007). Contingency tables summarizing the number of correct and incorrect responses to each task in the original vs. reference data are available in Appendix 7.

The selection of an indicator for comparison between the original and reference responses (Step 3) entailed careful examination of the procedure's null hypothesis in case of the 2PL IRT model. For each task, the null hypothesis states that the same set of parameters underlies the responses to the task at all time points; if the null hypothesis is true and the model fits the data, then the parameters estimated in Step 2 are approximations of the true

non-variant parameters. In such case, the original and reference responses compose two response patterns which are associated with very similar sets of parameters (the true and estimated parameters, respectively). In this null case, even with negligible differences between the true and estimated parameters, the original and reference responses are still not expected to be identical due to the nature of the 2PL IRT model: since the *probability* of a correct response is modeled as a function of the test taker's ability, the two sets of responses may differ due to chance. Moreover, the differences between the original and reference responses are expected to vary along the continuum of test taker ability, as demonstrated in Figure 1.

Figure 1

Item characteristic curve depicting the probability of responding correctly to a task according to the 2PL IRT model (a = 1.5*, b* = -1*)*

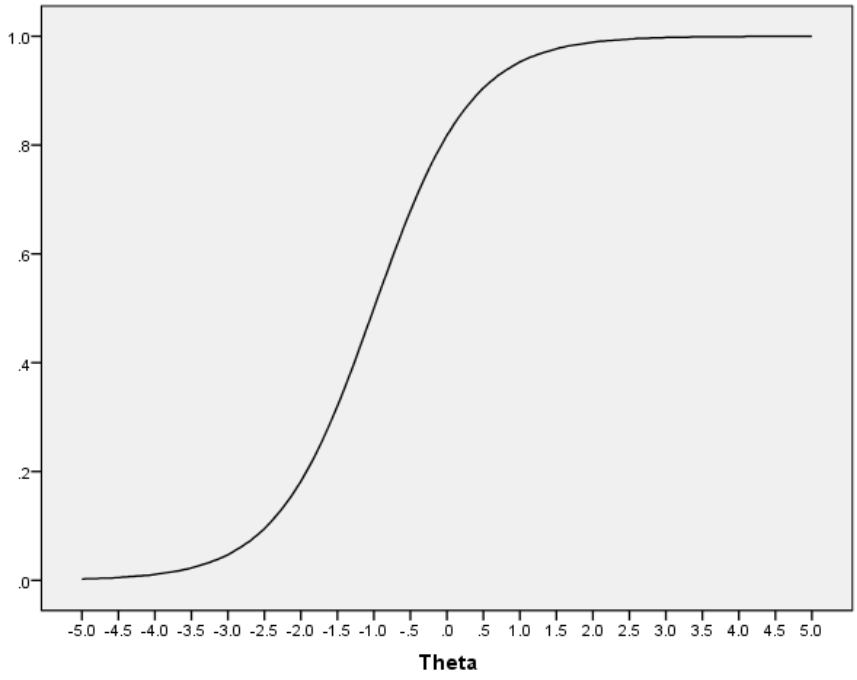


Figure 1 depicts an ICC relating test taker ability (theta) to the probability of responding correctly to a task in accordance with the 2PL IRT model. The smallest variation among responses associated with similar abilities is expected for test takers whose probability of responding correctly is either close to zero or close to 1, i.e. for test takers with either very low or very high ability in relation to the task's difficulty parameter. Moreover, the steeper the ICC at its inflection point (i.e., the larger the task's discrimination parameter), the smaller the interval around the difficulty parameter for which high variation is expected among the responses. In summary, the characteristics of the 2PL IRT model dictate that the larger the number of test takers with abilities in proximity to the task's difficulty parameter, the larger the overall inconsistency between two sets of possible responses for the entire sample; this “base variation” for IPD-free tasks depends on the magnitude of the discrimination parameter as well as on the task's location (difficulty) in relation to the distribution of test taker abilities. Ideally, IPD should be evaluated over and above this base variation stemming from the probabilistic nature of the model. However, the goal of using decision rules which take the base variation into account seems unrealistic in practice, since the true parameters upon which this variation depends are generally unknown.

An overall evaluation of the discrepancy between the original and reference responses to a single task relies upon a 2X2 contingency table of correct and incorrect responses in the two sets of data. A generic form of this table is presented in Table 12. As noted earlier, the original and reference responses are *paired* in the sense that each reference response corresponds to a specific original response, with both responses linked to the same test taker. Therefore, using the notation provided in Table 12, the interest lies in the values of the inner

Table 12

A contingency table of original and reference dichotomous responses to a task

Original Response	**Reference Response**		
	Correct	Incorrect	Total
Correct	N_{11}	N_{10}	$N_{11}+N_{10}$
Incorrect	N_{01}	N_{00}	$N_{01}+N_{00}$
Total	$N_{11}+N_{01}$	$N_{10}+N_{00}$	$N_{11}+N_{10}+N_{01}+N_{00}$

cells: N_{00} , N_{01} , N_{10} , and N_{11} , rather than the margins which represent the total number of correct and incorrect responses in the two sets of data. Inconsistencies between the original and reference responses are reflected in the values of N_{01} – the number of correct reference responses corresponding to incorrect original responses, and N_{10} – the number of incorrect reference responses corresponding to original responses which were correct; under the null hypothesis, the original and reference responses are based on the same non-drifting parameters, and therefore any inconsistency originates in chance alone. In other words, under the null hypothesis, the test takers' probability of responding correctly in the reference data is the same as their probability of responding correctly in the original data. Therefore, in the notation of Table 12, in a hypothetical scenario of all test takers having the same ability, the procedure's null hypothesis corresponds to

$$\frac{N_{11}+N_{01}}{N_{11}+N_{10}+N_{01}+N_{00}}=\frac{N_{11}+N_{10}}{N_{11}+N_{10}+N_{01}+N_{00}} \quad ,$$

which is the null hypothesis tested by the McNemar test (McNemar, 1947) of equality of

proportions for non-independent samples which are based on matched individuals. Readers should note that the equation above is equivalent to

$$N_{11} + N_{01} = N_{11} + N_{10} \quad ,$$

or

$$N_{01} = N_{10} \quad ;$$

therefore, the null hypothesis can be understood as equality of the two types of inconsistency between the original and reference responses. In the notation of Table 12, the McNemar test statistic is of the form

$$\frac{(N_{01} - N_{10})^2}{N_{01} + N_{10}} \quad ,$$

which, for a large sample size, follows a Chi-square distribution with one degree of freedom (McNemar, 1947). In the current analysis, the McNemar statistic was used for evaluating the equality of proportions of correct responses within the original and reference data for the entire sample, with a statistically significant difference taken as an indication of drift.

Table 13 shows the percentage of correct responses in the original and reference data for each task, and the results of the McNemar test for all tasks under both conditions (with and without changes in the test taker ability distribution over time). Reader should note that equality of proportions does not imply lack of inconsistency between the original and reference responses. For example, the percentage of correct responses to Task 18 without changes in the ability distribution over time was 53.4 and 53.5 in the original and reference data, respectively (Table 13); of a sample size of 150,000, this task had 32,516 correct

responses in the original data which were associated with incorrect responses in the reference data, and 32,629 incorrect responses in the original data paired with correct responses in the reference data (see Appendix 7).

Table 13

Percentage of correct responses in the original and reference data, and results of the McNemar test for each task under both conditions (with and without changes in the test taker ability distribution over time)

Task	Parameter(s) Changing over Time	True parameters at Time 1		No Changes in Ability Distribution				Ability Distribution Changing			
				Percentage of Correct Responses		McNemar Test		Percentage of Correct Responses		McNemar Test	
		a	*b*	Orig.	Ref.	χ^2	*p*	Orig.	Ref.	χ^2	*p*
Task 1	-	1.44	-1.05	75.2	76.1	41.74	<.001	77.0	77.7	24.80	<.001
Task 2	-	1.77	0.83	27.6	27.0	20.40	<.001	30.2	29.6	20.43	<.001
Task 3	-	1.13	-0.14	53.3	53.5	1.62	.203	55.4	55.7	2.17	.141
Task 4	-	1.12	-0.22	55.0	55.1	0.47	.491	57.3	57.4	0.70	.402
Task 5	***a***	0.96	-1.04	66.7	66.8	0.30	.581	68.0	68.2	1.70	.192
Task 6	-	0.58	0.01	50.1	50.1	0.03	.865	51.2	51.2	0.06	.805
Task 7	-	1.07	-0.88	68.5	68.8	2.09	.148	70.4	70.9	10.59	.001
Task 8	-	0.93	-0.24	54.9	55.0	0.33	.564	56.8	56.9	0.45	.502
Task 9	-	1.20	-1.43	79.5	80.0	14.60	<.001	81.1	81.6	16.72	<.001
Task 10	-	1.92	-0.90	74.9	75.6	30.44	<.001	77.0	77.8	45.63	<.001
Task 11	-	0.57	-0.90	61.6	61.9	2.05	.153	62.9	63.1	0.88	.349
Task 12	***a*** , ***b***	0.48	0.01	51.4	51.5	0.53	.465	51.7	52.0	1.98	.160
Task 13	-	1.38	0.82	30.3	29.9	6.00	.014	32.5	32.3	2.87	.090
Task 14	-	1.11	-1.20	74.8	75.2	11.26	.001	76.5	77.2	24.01	<.001
Task 15	***b***	0.28	0.09	51.1	51.2	0.30	.585	52.7	52.5	1.75	.186
Task 16	-	1.36	1.22	22.3	21.8	14.50	<.001	24.1	23.3	32.45	<.001
Task 17	-	0.25	0.35	48.0	48.2	0.45	.500	48.7	48.7	0.02	.877
Task 18	-	0.82	-0.18	53.4	53.5	0.19	.661	55.0	55.1	0.20	.657
Task 19	-	1.33	0.04	49.0	49.3	3.63	.057	51.3	51.5	1.04	.309
Task 20	-	1.27	-0.08	52.2	52.4	1.52	.218	54.5	54.5	0.01	.933
Task 21	-	0.96	1.17	28.0	28.0	0.25	.614	29.7	29.4	4.48	.034
Task 22	-	0.63	-0.88	62.6	62.6	0.00	.994	63.6	63.8	1.69	.193
Task 23	***a***	1.25	1.85	16.8	16.3	19.10	<.001	18.1	17.7	9.11	.003
Task 24	-	1.52	-0.44	61.6	62.0	9.63	.002	64.0	64.3	4.99	.025
Task 25	-	1.06	-1.19	73.5	74.0	8.76	.003	75.5	76.0	10.82	.001
Task 26	***b***	0.63	-1.09	70.3	70.4	0.80	.370	71.5	71.7	1.87	.172
Task 27	-	0.87	0.08	48.6	48.4	2.46	.117	50.5	50.4	0.32	.572
Task 28	***a*** , ***b***	1.95	-0.97	82.8	83.3	23.17	<.001	83.9	84.5	26.98	<.001
Task 29	-	0.48	0.53	44.1	43.8	2.28	.131	45.2	45.1	0.09	.763
Task 30	-	2.61	0.94	21.8	20.9	74.91	<.001	24.1	23.3	65.01	<.001

Note. For all McNemar tests, N = 150,000 and *df* = 1.
a = Discrimination parameter, *b* = Difficulty parameter

Table 13 indicates that for all tasks under both conditions, the percentage of correct responses in the original and reference data was similar, with the largest absolute difference equal to 0.9% (Task 1 and Task 30 without changes in the ability distribution over time). The percentage of correct responses to a task was consistently higher when the ability distribution was changing over time, with an average difference of 1.68% and 1.69% for the original and reference responses, respectively. This result was expected since under the condition of changes in the ability distribution, the average ability was modeled as rising over time.

For many of the tasks, the difference between the percentage of correct responses in the original and reference data was found statistically significant, without a clear association with the tasks undergoing true changes in their parameters over time. Under both conditions, only two of the six tasks with true drift were identified by the McNemar test: Task 23 (true change in discrimination; $\chi^2(1) = 19.10$, $p < .001$ without changes in the ability distribution over time; $\chi^2(1) = 9.11$, $p = .003$ with the ability distribution changing over time) and Task 28 (true changes in difficulty and discrimination; $\chi^2(1) = 23.17$, $p < .001$ without changes in the ability distribution over time; $\chi^2(1) = 26.98$, $p < .001$ with the ability distribution changing over time). Using a significance level of .01, under each of the two conditions (with and without changes in the test taker ability distribution over time) 9 of the 24 non-drifting tasks were falsely identified as exhibiting drift. The identification was similar across the two conditions, with a correlation of $r(30) = .91$ between the Chi-square values of the McNemar test with and without changes in the ability distribution over time.

In order to further investigate the lack of comparability between tasks with true drift and those identified as drifting by the McNemar test, two approaches were taken. First, since

the McNemar was initially applied here as an overall test which ignores the variation in test taker ability, a second round of analyses involved conducting separate McNemar tests along narrow intervals of the ability continuum. Selected results of these refined comparisons are shown in Appendix 8. The tasks selected for this demonstration were all six tasks with true drift (Tasks 5, 12, 15, 23, 26, and 28) along with two non-drifting tasks: one (Task 1) which was falsely identified as exhibiting drift in the overall McNemar test, and another (Task 3) which was correctly identified as non-drifting. The results indicate that an overall significant McNemar test may include intervals along the ability continuum for which the difference between proportions of correct responses in the original vs. reference data is not significant, and vice versa – when an overall McNemar test was not found significant, there may still be subgroups of test takers at specific ranges of abilities for which the difference is significant. In most cases, no specific pattern of significance (e.g., significant differences near the task's difficulty parameter, or significant differences towards one or both edges of the ability continuum) was identified.

As a second approach towards investigating the poor association between tasks with true drift and those identified as drifting by the overall McNemar test, a comparison was made between the observed results and those found under a null case of no drift in all tasks. For the purpose of simulating this null case, a new set of original data was created in the following way. Under each condition (with and without changes in the test taker ability distribution over time), responses at Time 2 and Time 3 were created for all tasks *assuming that the task parameters remained the same* at in Time 1 (i.e., no drift for all tasks). The proposed procedure was then applied to the null case in the same manner as described above.

The results of the proposed procedure previously shown in Table 13 – percentage of correct responses in the original and referent data, along with overall McNemar tests for all tasks – are presented in Table 14 in comparison to the results found for the null case of no drift.

First and foremost, the results presented in Table 14 indicate a high rate of false positives in the null case of no IPD – tasks flagged as showing drift when in fact, the true parameters remained the same at all three time points. Using a significance level of .01, the overall McNemar test resulted in 13 of the 30 tasks having significant Chi-square values under each condition (with and without changes in the test taker ability distribution over time) in the null case of no drift. The results were similar across the two conditions, with a correlation of $r(30) = .85$ between the Chi-square values of the McNemar test for the null case with and without changes in the ability distribution over time.

Second, the results presented in Table 14 show that in the current implementation of proposed procedure, the detection of drift when it was present in the data did not necessarily improve when analyzed in relation to the base (null) case of the same tasks without drift (i.e., tasks which are characterized by the same initial difficulty and discrimination parameters). Using a significance level of .01, for the condition of no changes in the test taker ability distribution over time (Table 14a), the identification of drift for the drifting tasks differed in only one case, and in the opposite direction: the overall McNemar test for Task 5 was significant in the null case of no drift, and not significant when the task's discrimination parameter was changing over time. When the test taker ability distribution was changing over time (Table 14b), the identification of drift for the drifting tasks differed in two cases, both in the opposite direction: the overall McNemar tests for Task 5 and Task 26 were significant in

Table 14

Results of the proposed procedure in comparison to a null case of all tasks maintaining the same parameters over time

a. No changes in test taker ability distribution over time

Task	Parameter(s) Changing over Time	True parameters at Time 1*		With IPD**				Null Case – No IPD			
				Percentage of Correct Responses		McNemar Test		Percentage of Correct Responses		McNemar Test	
		a	*b*	Orig.	Ref.	χ^2	*p*	Orig.	Ref.	χ^2	*p*
Task 1	-	1.44	-1.05	75.2	76.1	41.74	<.001	75.1	75.8	25.43	<.001
Task 2	-	1.77	0.83	27.6	27.0	20.40	<.001	27.7	26.9	32.20	<.001
Task 3	-	1.13	-0.14	53.3	53.5	1.62	.203	53.1	53.2	0.17	.680
Task 4	-	1.12	-0.22	55.0	55.1	0.47	.491	55.0	55.0	0.05	.827
Task 5	***a***	0.96	-1.04	66.7	66.8	0.30	.581	70.0	70.5	12.66	<.001
Task 6	-	0.58	0.01	50.1	50.1	0.03	.865	49.9	49.5	4.68	.031
Task 7	-	1.07	-0.88	68.5	68.8	2.09	.148	68.5	69.0	13.29	<.001
Task 8	-	0.93	-0.24	54.9	55.0	0.33	.564	54.7	54.7	0.26	.611
Task 9	-	1.20	-1.43	79.5	80.0	14.60	<.001	79.5	80.2	23.93	<.001
Task 10	-	1.92	-0.90	74.9	75.6	30.44	<.001	74.9	75.7	36.96	<.001
Task 11	-	0.57	-0.90	61.6	61.9	2.05	.153	61.9	62.0	0.47	.493
Task 12	***a*** , ***b***	0.48	0.01	51.4	51.5	0.53	.465	49.9	49.6	3.31	.069
Task 13	-	1.38	0.82	30.3	29.9	6.00	.014	30.2	29.6	20.67	<.001
Task 14	-	1.11	-1.20	74.8	75.2	11.26	.001	74.7	75.1	9.50	.002
Task 15	***b***	0.28	0.09	51.1	51.2	0.30	.585	49.6	49.3	1.82	.177
Task 16	-	1.36	1.22	22.3	21.8	14.50	<.001	22.3	21.7	19.20	<.001
Task 17	-	0.25	0.35	48.0	48.2	0.45	.500	47.8	47.9	0.08	.777
Task 18	-	0.82	-0.18	53.4	53.5	0.19	.661	53.3	53.7	4.08	.043
Task 19	-	1.33	0.04	49.0	49.3	3.63	.057	48.9	48.8	0.28	.597
Task 20	-	1.27	-0.08	52.2	52.4	1.52	.218	51.9	52.1	1.00	.317
Task 21	-	0.96	1.17	28.0	28.0	0.25	.614	27.8	27.6	1.44	.230
Task 22	-	0.63	-0.88	62.6	62.6	0.00	.994	62.4	62.5	0.42	.515
Task 23	***a***	1.25	1.85	16.8	16.3	19.10	<.001	13.8	13.2	22.98	<.001
Task 24	-	1.52	-0.44	61.6	62.0	9.63	.002	61.3	61.6	5.13	.024
Task 25	-	1.06	-1.19	73.5	74.0	8.76	.003	73.8	74.2	8.05	.005
Task 26	***b***	0.63	-1.09	70.3	70.4	0.80	.370	65.5	65.8	2.76	.097
Task 27	-	0.87	0.08	48.6	48.4	2.46	.117	48.6	48.6	0.06	.804
Task 28	***a*** , ***b***	1.95	-0.97	82.8	83.3	23.17	<.001	76.4	77.1	29.85	<.001
Task 29	-	0.48	0.53	44.1	43.8	2.28	.131	44.2	43.9	2.55	.110
Task 30	-	2.61	0.94	21.8	20.9	74.91	<.001	21.9	20.9	79.83	<.001

* In the null case of no IPD, the designated parameters were modeled as the true parameters at all time points.

** True IPD for highlighted tasks only.

Note. For all McNemar tests, N = 150,000 and *df* = 1.
a = Discrimination parameter, *b* = Difficulty parameter

(continued in the next page)

Table 14 (continued)

b. Test taker ability distribution changing over time

Task	Parameter(s) Changing over Time	True parameters at Time 1*		With IPD**				Null Case – No IPD			
				Percentage of Correct Responses		McNemar Test		Percentage of Correct Responses		McNemar Test	
		a	*b*	Orig.	Ref.	χ^2	*p*	Orig.	Ref.	χ^2	*p*
Task 1	-	1.44	-1.05	77.0	77.7	24.80	<.001	77.3	78.0	27.77	<.001
Task 2	-	1.77	0.83	30.2	29.6	20.43	<.001	30.2	29.6	19.69	<.001
Task 3	-	1.13	-0.14	55.4	55.7	2.17	.141	55.5	55.7	2.15	.143
Task 4	-	1.12	-0.22	57.3	57.4	0.70	.402	57.2	57.3	0.30	.585
Task 5	***a***	0.96	-1.04	68.0	68.2	1.70	.192	71.5	72.0	13.22	<.001
Task 6	-	0.58	0.01	51.2	51.2	0.06	.805	51.1	51.2	0.27	.601
Task 7	-	1.07	-0.88	70.4	70.9	10.59	.001	70.3	70.8	8.58	.003
Task 8	-	0.93	-0.24	56.8	56.9	0.45	.502	56.9	57.3	4.19	.041
Task 9	-	1.20	-1.43	81.1	81.6	16.72	<.001	81.0	81.5	11.71	.001
Task 10	-	1.92	-0.90	77.0	77.8	45.63	<.001	77.2	78.0	46.69	<.001
Task 11	-	0.57	-0.90	62.9	63.1	0.88	.349	62.9	63.0	0.21	.650
Task 12	***a*** , ***b***	0.48	0.01	51.7	52.0	1.98	.160	51.2	51.2	0.02	.896
Task 13	-	1.38	0.82	32.5	32.3	2.87	.090	32.5	32.0	13.29	<.001
Task 14	-	1.11	-1.20	76.5	77.2	24.01	<.001	76.4	76.7	4.13	.042
Task 15	***b***	0.28	0.09	52.7	52.5	1.75	.186	50.1	50.2	0.08	.774
Task 16	-	1.36	1.22	24.1	23.3	32.45	<.001	24.2	23.6	16.24	<.001
Task 17	-	0.25	0.35	48.7	48.7	0.02	.877	48.5	48.4	0.14	.708
Task 18	-	0.82	-0.18	55.0	55.1	0.20	.657	55.0	55.1	0.72	.397
Task 19	-	1.33	0.04	51.3	51.5	1.04	.309	51.5	51.5	0.00	.993
Task 20	-	1.27	-0.08	54.5	54.5	0.01	.933	54.4	54.4	0.15	.700
Task 21	-	0.96	1.17	29.7	29.4	4.48	.034	29.6	29.3	4.37	.037
Task 22	-	0.63	-0.88	63.6	63.8	1.69	.193	63.8	64.1	3.52	.061
Task 23	***a***	1.25	1.85	18.1	17.7	9.11	.003	15.1	14.5	22.94	<.001
Task 24	-	1.52	-0.44	64.0	64.3	4.99	.025	63.9	64.3	8.89	.003
Task 25	-	1.06	-1.19	75.5	76.0	10.82	.001	75.5	75.8	5.28	.022
Task 26	***b***	0.63	-1.09	71.5	71.7	1.87	.172	66.6	67.2	13.76	<.001
Task 27	-	0.87	0.08	50.5	50.4	0.32	.572	50.5	50.7	2.38	.123
Task 28	***a*** , ***b***	1.95	-0.97	83.9	84.5	26.98	<.001	78.8	79.6	46.33	<.001
Task 29	-	0.48	0.53	45.2	45.1	0.09	.763	45.4	45.2	0.57	.451
Task 30	-	2.61	0.94	24.1	23.3	65.01	<.001	24.3	23.6	41.80	<.001

* In the null case of no IPD, the designated parameters were modeled as the true parameters at all time points.
** True IPD for highlighted tasks only.

Note. 1. For all McNemar tests, N = 150,000 and $df = 1$.
2. a = Discrimination parameter, b = Difficulty parameter

the null case of no drift, and not significant when the task's discrimination (Task 5) and difficulty (Task 26) parameters were changing over time. Overall, the results of the McNemar test with and without true IPD were similar, with a correlation of $r(30) = .94$ between Chi-square values when the test taker ability distribution remained the same over time, and $r(30) = .83$ with changes in the ability distribution.

The failure of the two approaches to provide a reasonable explanation to the functionality of the proposed procedure calls for a conjecture that something other than drift lies in the basis of the McNemar test results. In searching for this factor, it is worth noting that the Chi-square values were found to be related to the tasks' discrimination parameters (at Time 1): $r(30) = .81$ both with and without changes in the test taker ability distribution over time (the correlation for the null case of no IPD was $r(30) = .82$ without changes in the ability distribution, and $r(30) = .79$ with the ability distribution changing). This result signifies that the larger the discrimination parameter, the higher the overall discrepancy between the two types of mismatched responses: correct original responses which are paired with incorrect reference responses, and incorrect original responses corresponding to correct reference responses. In addition, despite attempts to keep the estimated abilities on the same scale as the true abilities upon which the original simulation was based, the statistics of test taker ability provided in Appendix 6 suggest that there may have been a difference in the abilities' scaling. When the test taker ability distribution remained the same over time, the estimated abilities ranged from -2.80 to 2.54, in comparison to a range of -4.64 to 4.42 for the true abilities; with the ability distribution changing, the estimated abilities were between -2.85 and 2.47, while the true abilities ranged between -4.64 to 4.88 (Appendix 6). The mean absolute difference between true and estimated abilities was $M = 0.31$ ($SD = 0.24$) and $M = 0.30$ ($SD = 0.23$) with and without changes in the test taker ability distribution over time, respectively, and the correlation between true and estimated abilities was $r(150{,}000) = .925$ for both conditions (see scatter plots of estimated vs. true ability in Appendix 9). This information suggests that scaling, rather than estimation error, may be a central factor

explaining the differences between the true and estimated abilities. A scaling difference suggests that the original and reference responses cannot be strictly considered matched; this result undermines the selection of the McNemar test as an indicator for comparison between the two sets of responses.

Summary of the Results Obtained by the Three IPD Analyses

The conclusions regarding drift for all tasks under both conditions (with and without changes in the test taker ability distribution over time) using the MGDIF and LogR methods are summarized in Table 15. A detection of drift which is driven by p-values smaller than .01 (Table 15a) resulted in all tasks with true drift identified as drifting by both methods (with some cases of incorrect specification of the type of drift using LogR); however, the number of non-drifting tasks falsely identified as exhibiting drift was extremely high for the MGDIF method (22/24 under each of the two conditions), and somewhat lower (but nonetheless high) for LogR (8/24 and 14/24 with and without changes in the ability distribution over time, respectively). The two tasks which were not identified as drifting by the MGDIF method (Task 6 and Task 17 under both conditions), were also not identified with drift in the LogR analysis. The detection of drift according to extreme Chi-square values (Table 15b) resulted in complete agreement between the MGDIF and LogR results, in which all tasks with true drift, and only those tasks, were identified as drifting under both conditions (with some cases of incorrect specification of the type of drift using LogR). To summarize, the two known methods applied in the context of this demonstration produced results which were valid and similar in essence.

The procedure proposed in this paper, however, resulted in identification of drift which did not align with the true drift modeled in the data. Moreover, when the procedure was applied on a null case in which none of the tasks was undergoing true changes in

Table 15

Detection of drift according to the MGDIF and LogR methods

a. Identification of drift according to traditional hypothesis testing (significance level α=.01)

Task	Parameter(s) changing over time	True parameters at Time 1		No changes in ability distribution		Ability distribution changing	
		a	***b***	**MGDIF**	**LogR**	**MGDIF**	**LogR**
Task 1	-	1.44	-1.05	Drift	Drift (uniform)	Drift	-
Task 2	-	1.77	0.83	Drift	Drift (overall)	Drift	Drift (non-uniform; a)
Task 3	-	1.13	-0.14	Drift	-	Drift	-
Task 4	-	1.12	-0.22	Drift	Drift (uniform)	Drift	Drift (non-uniform; a)
Task 5	***a***	0.96	-1.04	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a,b)
Task 6	-	0.58	0.01	-	-	-	-
Task 7	-	1.07	-0.88	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a)
Task 8	-	0.93	-0.24	Drift	Drift (uniform)	Drift	-
Task 9	-	1.20	-1.43	Drift	Drift (uniform)	Drift	-
Task 10	-	1.92	-0.90	Drift	Drift (uniform)	Drift	-
Task 11	-	0.57	-0.90	Drift	Drift (uniform)	Drift	-
Task 12	***a*** , ***b***	0.48	0.01	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a,b)
Task 13	-	1.38	0.82	Drift	-	Drift	Drift (non-uniform; a)
Task 14	-	1.11	-1.20	Drift	Drift (uniform)	Drift	-
Task 15	***b***	0.28	0.09	Drift	Drift (uniform)	Drift	Drift (uniform)
Task 16	-	1.36	1.22	Drift	-	Drift	Drift (non-uniform; a)
Task 17	-	0.25	0.35	-	-	-	-
Task 18	-	0.82	-0.18	Drift	Drift (uniform)	Drift	Drift (non-uniform; a)
Task 19	-	1.33	0.04	Drift	-	Drift	Drift (non-uniform; a)
Task 20	-	1.27	-0.08	Drift	-	Drift	-
Task 21	-	0.96	1.17	Drift	-	Drift	-
Task 22	-	0.63	-0.88	Drift	Drift (uniform)	Drift	-
Task 23	***a***	1.25	1.85	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a,b)
Task 24	-	1.52	-0.44	Drift	Drift (uniform)	Drift	-
Task 25	-	1.06	-1.19	Drift	Drift (uniform)	Drift	-
Task 26	***b***	0.63	-1.09	Drift	Drift (uniform)	Drift	Drift (non-uniform; a,b)
Task 27	-	0.87	0.08	Drift	-	Drift	-
Task 28	***a*** , ***b***	1.95	-0.97	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a,b)
Task 29	-	0.48	0.53	Drift	Drift (uniform)	Drift	-
Task 30	-	2.61	0.94	Drift	-	Drift	Drift (non-uniform; a)

Note: *a* = Discrimination parameter

b = Difficulty parameter

(continued in the next page)

Table 15 (continued)

b. Identification of drift according to outlier Chi-square values (among the analyzed tasks)

Task	Parameter(s) changing over time	True parameters at Time 1		No changes in ability distribution		Ability distribution changing	
		a	*b*	MGDIF	LogR	MGDIF	LogR
Task 1	-	1.44	-1.05	-	-	-	-
Task 2	-	1.77	0.83	-	-	-	-
Task 3	-	1.13	-0.14	-	-	-	-
Task 4	-	1.12	-0.22	-	-	-	-
Task 5	***a***	0.96	-1.04	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a,b)
Task 6	-	0.58	0.01	-	-	-	-
Task 7	-	1.07	-0.88	-	-	-	-
Task 8	-	0.93	-0.24	-	-	-	-
Task 9	-	1.20	-1.43	-	-	-	-
Task 10	-	1.92	-0.90	-	-	-	-
Task 11	-	0.57	-0.90	-	-	-	-
Task 12	***a*** **,** ***b***	0.48	0.01	Drift	Drift (non-uniform; a)	Drift	Drift (non-uniform; a,b)
Task 13	-	1.38	0.82	-	-	-	-
Task 14	-	1.11	-1.20	-	-	-	-
Task 15	***b***	0.28	0.09	Drift	Drift (uniform)	Drift	Drift (uniform)
Task 16	-	1.36	1.22	-	-	-	-
Task 17	-	0.25	0.35	-	-	-	-
Task 18	-	0.82	-0.18	-	-	-	-
Task 19	-	1.33	0.04	-	-	-	-
Task 20	-	1.27	-0.08	-	-	-	-
Task 21	-	0.96	1.17	-	-	-	-
Task 22	-	0.63	-0.88	-	-	-	-
Task 23	***a***	1.25	1.85	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a,b)
Task 24	-	1.52	-0.44	-	-	-	-
Task 25	-	1.06	-1.19	-	-	-	-
Task 26	***b***	0.63	-1.09	Drift	Drift (uniform)	Drift	Drift (uniform)
Task 27	-	0.87	0.08	-	-	-	-
Task 28	***a*** **,** ***b***	1.95	-0.97	Drift	Drift (non-uniform; a,b)	Drift	Drift (non-uniform; a,b)
Task 29	-	0.48	0.53	-	-	-	-
Task 30	-	2.61	0.94	-	-	-	-

Note: *a* = Discrimination parameter

b = Difficulty parameter

parameters over time, nearly half of the tasks (13/30 under each of the two condition) were still identified as drifting. The true discrimination parameters at Time 1 were found to have high correlation with the Chi-square values of the McNemar test. At this point, it is still unclear whether the results indicate a flaw in the procedure itself, or in the way it was implemented for the specific analysis presented in this paper. A more elaborate discussion around this question is provided ahead.

Discussion and Limitations

The two known methods applied for the detection of drift in the current simulation study – MGDIF and LogR – have had similar results in terms of accuracy in the detection of tasks with true drift. This outcome is interesting, since the null hypotheses tested by the two methods are different in nature. For each task, the MGDIF method employes an IRT-based procedure which tests the equality of both the difficulty and discrimination parameters across all groups – i.e., time points in the context of drift (Kim et al., 1995, p. 263). The LogR method uses an observed total score (rather than a latent factor) as a matching variable (in the current case, a total score based on all tasks except the task under study), and tests the equality of R-squared values of two nested models (Zumbo, 1999, p. 26) – Model 2 vs. Model 1, Model 3 vs. Models 2, and Model 3 vs. Model 1 in the current application – see description earlier in this paper.

The link between this use of logistic regression for detecting differential functioning and the MGDIF method can be established using the terms *uniform* and *non-uniform DIF*, which are depicted in both IRT-based and LogR terminology by Zumbo (1999). Uniform DIF, which is characterized in IRT-based terminology by differences in the difficulty parameter, is tested in the LogR method as a significant difference between the percentage of variance in the dependent variable (logit of the task responses) which is explained by Model 2 (in the current case, matching variable and time point) in comparison with Model 1 (matching variable alone). Non-uniform DIF, depicted in IRT terms as differences in the discrimination parameter, is characterized in the LogR method as a significant difference

between the percentage of variance in the dependent variable which is explained by Model 3 (matching variable, time point, and the interaction between the two) compared to Model 2. The MGDIF test can then, in principle, be perceived as parallel to the difference between Model 3 and Model 1 – a simultaneous test of uniform and non-uniform differential functioning (Swaminathan & Rogers, 1990, p. 364-365), which corresponds to differences in either difficulty or discrimination.

A careful examination of the MGDIF and LogR results with a detection of drift which is based on outlier Chi-square values (Table 15b and Table 11 for MGDIF and LogR, respectively) reveals that the MGDIF test and the test of difference between Model 3 and Model 1 in the LogR analysis were indeed in complete agreement; the statistics for the two tests – both evaluated against a Chi-square distribution with 4 degrees of freedom – were highly correlated ($r(30) = .92$ and $r(30) = .95$ with and without changes in the test taker ability distribution, respectively).

The LogR results provided a more refined description than the MGDIF method regarding the type of identified drift; however, the conclusions were not accurate in all cases.

False detection of uniform drift. Under both conditions (with and without changes in the test taker ability over time), in both cases of true changes only in discrimination (Task 5 and Task 23), uniform drift was identified by the LogR method in addition to non-uniform drift. Readers should note that in the 2PL IRT model, a decrease in the discrimination parameter of a task is characterized by the two following features: (1) the probability of responding correctly to the task increases for test takers whose abilities are lower than the task's difficulty parameter, and (2) the probability of responding correctly decreases for test

takers with abilities which are higher than the task's difficulty parameter. In light of these two attributes, the false detection of uniform drift for Task 5 and Task 23 could be explained by the fact that both tasks had a difficulty parameter which was far from the center of the ability distribution. Task 5 had a true difficulty parameter of b = -1.04, which is about one standard deviation lower than the mean true ability of test takers at Time 1; this means that the number of test takers at Time 2 and Time 3 who had a lower probability of responding correctly to the task (compared to test takers with the same ability at Time 1) was larger than the number of test takers at these time points who had a higher probability of responding correctly, a situation which could lead to a false conclusion of an increase in the task's difficulty parameter. In this case, the lack of balance was expected to be more extreme under the condition of changes in the test taker ability distribution over time, since the changes were modeled as an increase in the mean test taker ability – which resulted in an increased distance between the center of the ability distribution and the task's difficulty parameter. An opposite situation has likely occurred in the case of Task 23, which had a true difficulty parameter of b = 1.85; the lack of balance in this case could result in false identification of a decrease in the task's difficulty parameter.

Uniform drift which was not detected. In one case of true changes in both difficulty and discrimination – Task 12 without changes in the test taker ability distribution over time – only non-uniform drift was detected by the LogR method. It could be argued that a finding of non-uniform drift should be interpreted as an overall judgment which may include uniform drift; however, in the three other cases of changes in both difficulty and discrimination, both uniform and non-uniform drift were detected by the LogR method. The

uniform drift which was not detected in the current case could be a matter of judgment, since the Chi-square value of the uniform drift test was the highest of all values which were not flagged as extreme.

The implementation of the new procedure proposed in this paper did not seem to provide satisfactory results in identifying tasks with true drift. There may be several reasons to this outcome. First, in theory, the indicator selected for the comparison between original and reference responses – the McNemar test – should be used with groups of test takers which are matched on ability; it is possible that even with the abilities grouped into segments of 0.5 on the theta scale, this test was still too broad to provide meaningful results, especially around a task's difficulty parameter where the probability of a correct response may change fundamentally within an interval of 0.5. The McNemar test may function better when applied on narrower intervals on the ability continuum, or a different indicator might have provided better results. Indicators which could be considered in the future include the Phi correlation coefficient and odds of agreement (the proportion of responses which are identical in the original and reference data divided by the proportion of responses which are different). Both indicators use the overall 2X2 contingency table for each task (Table 12) rather than a more refined comparison based on ability; however, as mentioned earlier, a comparison which is based on ability is limited to the estimated abilities, as the true test taker parameters are unknown in practical testing frameworks.

The second possible explanation to the poor performance of the proposed procedure has to do with scaling. The procedure's second step (see Table 6) involves the creation of reference responses based on the parameters estimated in Step 1. In the case of the 2PL IRT

model, the original and reference responses to a non-drifting task with good model fit may still differ due to two reasons: the probabilistic nature of the model, and differences between the true and estimated model parameters. In order to minimize inconsistencies which are based on the latter, it is crucial that the true and estimated parameters will be placed on the same scale. This is true for all three model parameters: discrimination, difficulty, and test taker ability. Different scalings would result in a comparison between the original and reference responses which is meaningless, since the probability of a correct response is determined based on the specific values of the model parameters. In the current case, the statistics of test takers' estimated and true ability (Appendix 6) indicate that despite deliberate attempts to place both sets of parameters on the same scale, the estimated parameters resulted in a minimum and maximum which are different from those of the true parameters. This issue of scaling equivalence may be less influential in real data where the original responses are not strictly produced based on the applied model, and further studies are required in order to assess the functionality of the proposed procedure with real data.

The comparability of scalings raises a more general caveat regarding the proposed procedure – its dependence on the applied model. The procedure is not model-based in the sense that it can be applied with various models; however, once a model is selected, the procedure's implementation – parameter estimation and creation of reference responses which are free of drift – is highly based on the model's features. Even the comparison between original and reference data may entail using the model's characteristics in order to apply formal hypothesis testing. While the dependency on the selected model can be considered a shortcoming of the suggested procedure, it also highlights the fact that the

analysis presented in this paper represents only one possible application of the proposed procedure in relation to one specific model. The procedure may still prove useful if implemented using different models, and the current failure to produce adequate results does not indicate that the general process should be disregarded.

Although all steps of the proposed procedure (Table 6) seem to follow a logical line of thought, two notes regarding Step 2 are in place. First, this step is currently stated as taking the approach of simulating a single set of reference responses. This strategy may be sufficient when the sample size is large enough to be considered analogous to a population (as in the current application). With smaller numbers of test takers, however, it might be more adequate to repeat this step multiple times, so the original responses could be compared to a space of possible response patterns without drift rather than a single set of reference responses. Under this modification, each original response would have be compared to a series of reference responses, and a suitable indicator for such comparison would have to be selected in Step 3. Despite the large sample size, the current implementation of the proposed procedure could have benefited from such approach, and the results could have been different in terms of accuracy in the detection of tasks with true drift.

A second comment regarding Step 2 challenges the entire notion of simulating reference responses, in the following way. In cases of IRT models, including the current analysis in which the 2PL model was applied, instead of simulating reference responses – the original responses could be referred directly to the ICCs which derive from the parameters estimated in Step 1, using the estimated ability as a base for determining the probability of a correct response for each test taker under the assumption of no drift. This option has the

potential of using the full information derived from Step 1 directly, rather than a projection in the form of a single – or multiple – reference response(s). While such modification may eliminate one of the unique features of the proposed procedure – the comparison between responses – it should be kept in mind that the non-parametric appearance of this part of the procedure is misleading, since the reference responses are simulated based on the parameters estimated in the first step. For example, in the current case of the 2PL IRT model, the applied procedure may seem to be distinguished from existing methods in which ICCs are compared (Raju, 1988; Kim & Cohen, 1991; Wells et al., 2014), but in fact, the reference responses (and the original responses, in the current case of a simulation study) are created based on ICCs.

In addition to the two comments regarding Step 2 of the proposed procedure, it should be noted that although the procedure is presented as a general approach to the detection of drift which may be implemented in various ways, it was still formulated with IRT-like framework in mind, and its application with different types of measurement models is yet to be investigated.

Using each of the three methods, no essential differences were found in the detection of drift under the two conditions – with and without changes in the test taker ability distribution over time. This finding suggests that given the features selected for the simulated data, the methods were robust to changes in the test taker ability distribution over time, when the changes were limited to mean differences of 0.1 standard deviations between each pair of consecutive time points. A larger change in the mean ability over time, or an additional change in variance, could have resulted in a different outcome. In the future, it would be

interesting to test whether the current equivalence between the results of the MGDIF and LogR methods holds under different extents of instability of the test taker ability distribution over time. Another explanation to the similarity between the results under the two condition relates to the fact that the same group of test takers was used at Time 1 under both conditions (see Table 8). That is, the analyses under the two conditions were not independent in the sense that for each task, the first 50,000 responses – a third of the data set – were identical for both conditions. This design was selected in order to minimized differences between the results under both conditions which could be caused by chance alone; however, since the two analyses resulted in similar outcomes, it would be interesting to assess whether this similarity holds when two different groups of test takers (with the same mean and variance of test taker ability) are used at Time 1.

The analyses presented in this paper are also limited in the sense that no iterative process – removal of tasks identified with drift and re-examination of the remaining tasks – was conducted. Such processes are important since in all three methods, the test taker abilities are represented using some factor (latent ability, corrected total score) which is derived from the data regarding all tasks – including those identified with drift. The values of the test taker ability parameters influence the entire analysis, and a final conclusion of no drift should in fact be made using a data set which is free of concerns about drift. In addition, none of the three analyses presented in this paper included further examination of tasks identified as drifting in order to assess the nature of changes over time. The LogR method was the most informative in this sense, since it provided further indication of whether the differences between time points were uniform of non-uniform; however, even a conclusion of

this type might be insufficient in practical settings, where researchers may be interested in evaluating the size and direction of the changes as well as the part(s) of the ability continuum (e.g., test takers with lower abilities) in which the differences were most influential.

Another limitation of the study relates to the sample size. The large number of test takers (N=50,000 at each time point) was selected for the purpose of establishing stable results in the current implementation of the proposed procedure, with each original response being compared to a single reference response. The MGDIF and LogR methods did not require such a large sample size, but the same data set was used across all three analyses in order to establish a reliable comparison between the results of the two known methods and those of the proposed procedure. The large sample size resulted in most of the tasks identified as drifting using the MGDIF and LogR methods, and the results had to be evaluated using decision rules which are based on the magnitude of the Chi-square values rather than traditional p-values. In the future, it would be beneficial to compare the results of these two methods using a smaller a sample size, non-inflated Chi-square values, and a more objective conclusion about which tasks exhibit drift using p-values. Future simulation-based research should also use replications, so that the current analyses could be expanded to an evaluation of the Type I error rate and power in the detection of drift using the MGDIF method, LogR, and the proposed procedure.

References

Arce-Ferrer, A. J. & Bulut, O. (2016). Investigating Separate and Concurrent Approaches for Item Parameter Drift in 3PL Item Response Theory Equating. *International Journal of Testing, 0*, 1-22.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*(4), 610-619.

Cuevas, M. & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Mathématiques et Sciences Humaines. Mathematics and Social Sciences, 2012*(3), 45-59.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*(3), 265-300.

Donoghue, J. R. & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, *22*(1), 33-51.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457-459.

Hidalgo, M. D. & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS Program Statistics Research Technical Report No. 85-64). Princeton, NJ, USA: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1986). *Differential item performance and the Mantel-Haenszel procedure* (ETS Research Report No. RR-86-31). Princeton, NJ, USA: Educational Testing Service.

Kim, S-H., & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied Psychological Measurement*, *15*(3), 269-278.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*(3), 261-276.

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

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847-862.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157.

Meyer, J. P. (2014). *Applied Measurement with jMetrik*. New York, NY: Routledge.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrica*, *53*(4), 495-502.

Raju, N. S. (1990). Determining the significance of signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17*(1), 1-10.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Veerkamp, W. J. J. & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25*(4), 373-389.

Wells, C. S., Hambleton, R. K., Kirkpatrick, R., & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education, 27*(3), 214-231.

Woods, C., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532-547.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (Research Report No. RR-12-08). Princeton, NJ, USA: Educational Testing Service.

Appendices

Appendix 1

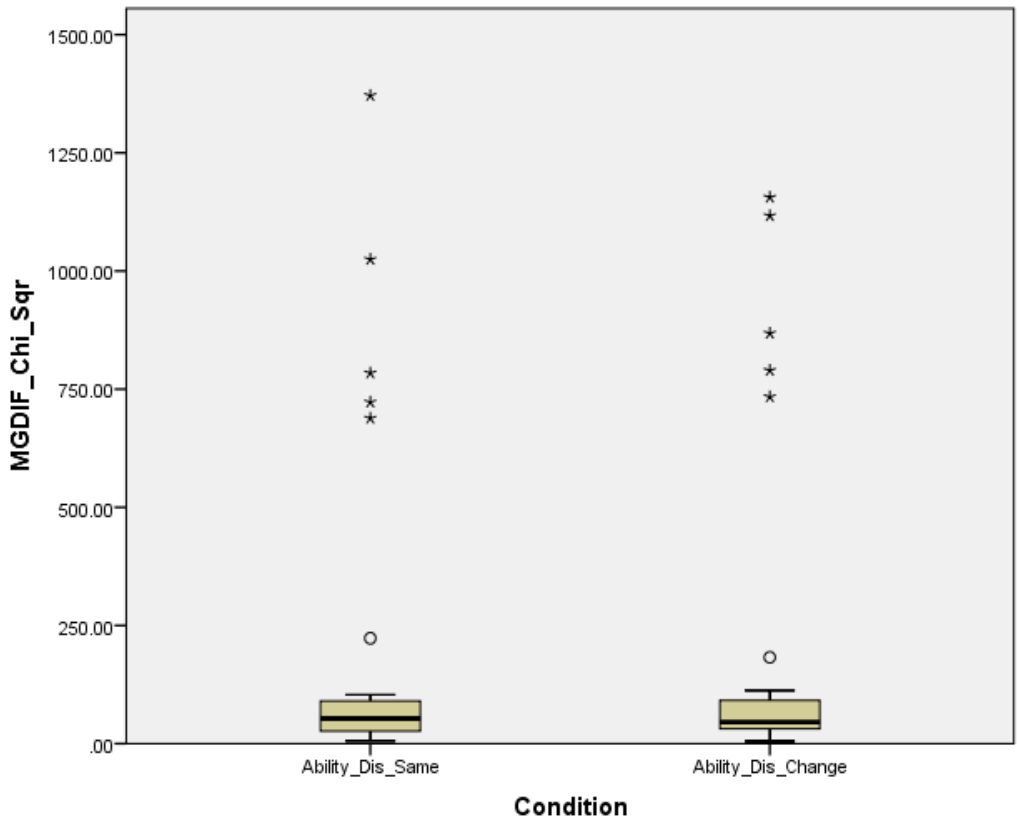
R code for the detection of IPD based on the Multiple-Groups DIF (MGDIF) method presented by Kim et al. (1995)

```
Thesis_data_dis_same = read.csv("Thesis data dis same _timePoint.csv",TRUE,",")
ConMat = contrastMatrix(2,"2PL")
difGenLord(Thesis_data_dis_same, "Time", c("Time2","Time3"), "2PL", c = NULL,
        engine = "ltm", nrFocal = 2, nrIter = 10, save.output = TRUE,
        output = c("Output KCP dis same", "default"))
```

Note: 1. The code refers to the condition of no changes in ability distribution over time. The code for the second condition (with changes in ability distribution) differs only by the identification of the input file (first command).
2. The R package 'difR' is required.

Appendix 2

Distributions of MGDIF Chi-Square values for the 30 simulated tasks under both conditions (with and without changes in the test taker ability distribution over time)



Note: The analysis of the variables' distributions and the resulting figure were produced using SPSS statistical software.
IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp

Appendix 3

SPSS syntax for the detection of IPD using Logistic Regression

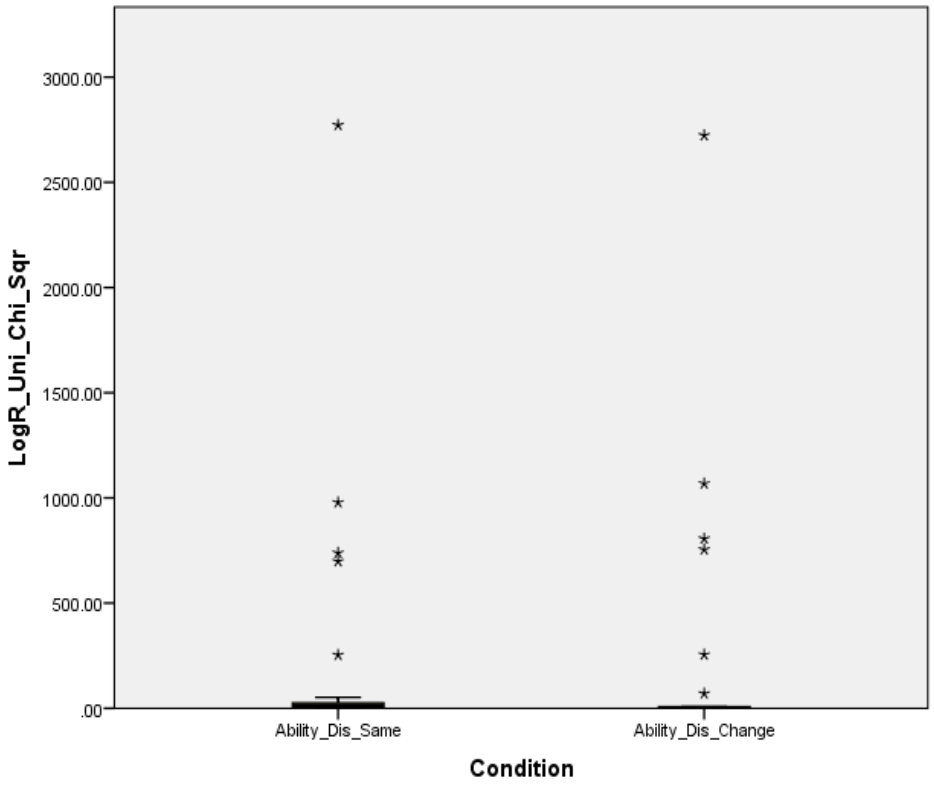
```
compute Task_res = Task1_res.
compute Tot_exc_task = Total_exc_task1.

* Uniform Drift.
LOGISTIC REGRESSION VARIABLES Task_res
  /METHOD=ENTER Tot_exc_task
  /METHOD=ENTER Time
  /CONTRAST (Time)=Indicator(1)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

* Non- Uniform Drift -- 2df test (interaction terms only).
LOGISTIC REGRESSION VARIABLES Task_res
  /METHOD=ENTER Tot_exc_task Time
  /METHOD=ENTER Time*Tot_exc_task
  /CONTRAST (Time)=Indicator(1)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

* Non- Uniform Drift -- 4df test (main effect and interaction together).
LOGISTIC REGRESSION VARIABLES Task_res
  /METHOD=ENTER Tot_exc_task
  /METHOD=ENTER Time Time*Tot_exc_task
  /CONTRAST (Time)=Indicator(1)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Note: The code refers to the identification of drift for a single task (Task 1).

Appendix 4

Detection of IPD using LogR: Distributions of Chi-square values for differences between nested models for the 30 simulated tasks under both conditions (with and without changes in the test taker ability distribution over time)

a. Test of uniform drift: Difference between Model 2 and Model 1 (*df*=2)

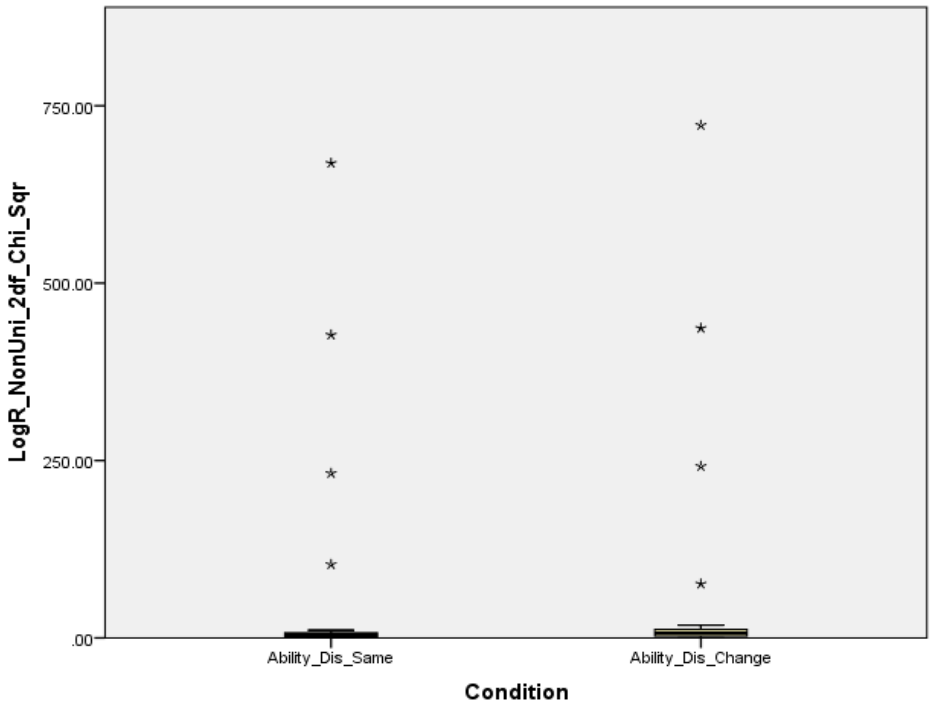


Note: 1. Variables used as predictors in the nested models:
Model 1. Rest score
Model 2. Rest score, Time point
Model 3. Rest score, Time point, interaction between Rest score and Time point
2. The analysis of the variables' distributions and the resulting figures were produced using SPSS statistical software.
IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp

(continued in the next page)

Appendix 4 (continued)

b. Test of non-uniform drift: Difference between Model 3 and Model 2 (*df*=2)

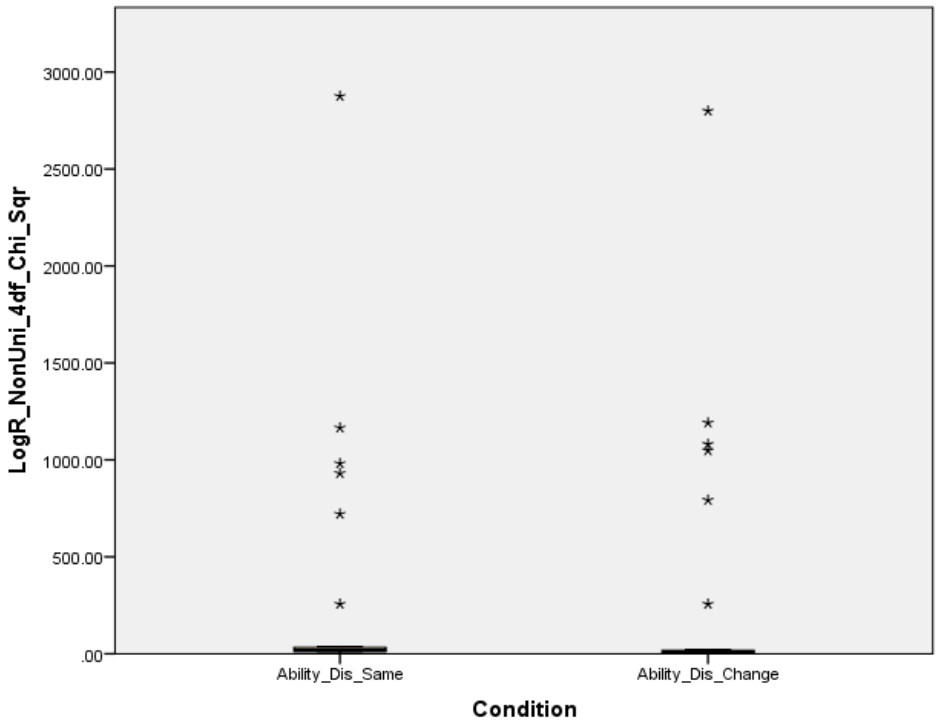


Note: 1. Variables used as predictors in the nested models:
Model 1. Rest score
Model 2. Rest score, Time point
Model 3. Rest score, Time point, interaction between Rest score and Time point
2. The analysis of the variables' distributions and the resulting figures were produced using SPSS statistical software.
IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp

(continued in the next page)

Appendix 4 (continued)

c. Test of Non-Uniform Drift: Difference between Model 3 and Model 1 (*df*=4)



Note: 1. Variables used as predictors in the nested models:
Model 1. Rest score
Model 2. Rest score, Time point
Model 3. Rest score, Time point, interaction between Rest score and Time point
2. The analysis of the variables' distributions and the resulting figures were produced using SPSS statistical software.
IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp

Appendix 5

Task parameters estimated using the proposed procedure for evaluating IPD, with and without changes in the test taker ability distribution over time

Task	Parameter(s) changing over time	True parameters at Time 1		Estimated Parameters			
				No changes in ability distribution		Ability distribution changing	
		a	*b*	*a*	*b*	*a*	*b*
Task 1	-	1.44	-1.05	1.42	-1.06	1.45	-1.14
Task 2	-	1.77	0.83	1.74	0.84	1.77	0.73
Task 3	-	1.13	-0.14	1.12	-0.15	1.14	-0.24
Task 4	-	1.12	-0.22	1.12	-0.22	1.12	-0.32
Task 5	***a***	0.96	-1.04	0.74	-1.04	0.73	-1.15
Task 6	-	0.58	0.01	0.58	0.00	0.58	-0.08
Task 7	-	1.07	-0.88	1.07	-0.89	1.08	-0.98
Task 8	-	0.93	-0.24	0.94	-0.24	0.93	-0.34
Task 9	-	1.20	-1.43	1.21	-1.41	1.20	-1.52
Task 10	-	1.92	-0.90	1.95	-0.89	1.95	-0.98
Task 11	-	0.57	-0.90	0.57	-0.89	0.57	-0.99
Task 12	***a*** , ***b***	0.48	0.01	0.28	-0.20	0.27	-0.25
Task 13	-	1.38	0.82	1.37	0.82	1.39	0.72
Task 14	-	1.11	-1.20	1.10	-1.21	1.12	-1.30
Task 15	***b***	0.28	0.09	0.28	-0.30	0.28	-0.39
Task 16	-	1.36	1.22	1.36	1.22	1.36	1.13
Task 17	-	0.25	0.35	0.25	0.32	0.26	0.21
Task 18	-	0.82	-0.18	0.81	-0.19	0.81	-0.28
Task 19	-	1.33	0.04	1.32	0.04	1.34	-0.05
Task 20	-	1.27	-0.08	1.28	-0.09	1.27	-0.18
Task 21	-	0.96	1.17	0.95	1.17	0.96	1.07
Task 22	-	0.63	-0.88	0.62	-0.90	0.63	-0.97
Task 23	***a***	1.25	1.85	1.04	1.84	1.03	1.75
Task 24	-	1.52	-0.44	1.51	-0.44	1.51	-0.53
Task 25	-	1.06	-1.19	1.05	-1.18	1.05	-1.30
Task 26	***b***	0.63	-1.09	0.62	-1.51	0.65	-1.54
Task 27	-	0.87	0.08	0.87	0.08	0.87	-0.02
Task 28	***a*** , ***b***	1.95	-0.97	1.71	-1.34	1.77	-1.38
Task 29	-	0.48	0.53	0.48	0.53	0.48	0.43
Task 30	-	2.61	0.94	2.60	0.94	2.62	0.85
	Mean	**1.10**	**-0.16**	**1.07**	**-0.21**	**1.07**	**-0.30**
	SD	**0.53**	**0.85**	**0.53**	**0.88**	**0.53**	**0.87**

Note: *a* = Discrimination parameter
b = Difficulty parameter

Appendix 6

Statistics of true test taker abilities and the abilities estimated using the proposed procedure for evaluating IPD

a. No changes in test taker ability distribution over time

Statistic	Overall (N=150,000)				Time 1 (N=50,000)		Time 2 (N=50,000)		Time 3 (N=50,000)	
	True θ	Estimated θ	Difference between Estθ and True θ	Absolute Difference between Estθ and True θ	True θ	Estimated θ	True θ	Estimated θ	True θ	Estimated θ
Mean	0.001	0.003	0.002	0.300	0.002	-0.021	0.003	0.007	-0.002	0.022
SD	0.999	0.926	0.379	0.231	1.003	0.948	0.998	0.923	0.996	0.905
Min	-4.645	-2.796	-2.273	0.000	-4.645	-2.796	-4.618	-2.796	-4.156	-2.796
Max	4.420	2.541	2.637	2.637	4.135	2.541	4.217	2.541	4.420	2.541

b. Test taker ability distribution changing over time

Statistic	Overall (N=150,000)				Time 1 (N=50,000)		Time 2 (N=50,000)		Time 3 (N=50,000)	
	True θ	Estimated θ	Difference between Estθ and True θ	Absolute Difference between Estθ and True θ	True θ	Estimated θ	True θ	Estimated θ	True θ	Estimated θ
Mean	0.101	0.004	-0.096	0.311	0.002	-0.101	0.105	0.012	0.196	0.103
SD	1.003	0.926	0.380	0.240	1.003	0.945	0.996	0.918	1.002	0.903
Min	-4.645	-2.853	-2.738	0.000	-4.645	-2.853	-3.976	-2.853	-3.880	-2.853
Max	4.877	2.475	1.968	2.738	4.135	2.475	4.216	2.475	4.877	2.475

Appendix 7

The number of correct and incorrect responses to each task in the original vs. reference data

Task	Parameter(s) changing over time	True parameters at Time 1		Original Response	No Changes in Ability Distribution		Ability Distribution Changing	
		a	*b*		Reference Response		Reference Response	
					Correct	Incorrect	Correct	Incorrect
1	-	1.44	-1.05	Correct	92,885	19,901	96,214	19,304
				Incorrect	21,212	16,002	20,296	14,186
2	-	1.77	0.83	Correct	21,207	20,258	24,252	20,996
				Incorrect	19,358	89,177	20,079	84,673
3	-	1.13	-0.14	Correct	50,708	29,304	54,109	29,051
				Incorrect	29,614	40,374	29,408	37,432
4	-	1.12	-0.22	Correct	53,106	29,325	56,931	28,948
				Incorrect	29,493	38,076	29,151	34,970
5	***a***	0.96	-1.04	Correct	70,317	29,786	72,838	29,145
				Incorrect	29,922	19,975	29,462	18,555
6	-	0.58	0.01	Correct	40,295	34,830	42,046	34,752
				Incorrect	34,784	40,091	34,686	38,516
7	-	1.07	-0.88	Correct	76,433	26,365	80,466	25,146
				Incorrect	26,699	20,503	25,882	18,506
8	-	0.93	-0.24	Correct	51,230	31,110	54,303	30,884
				Incorrect	31,255	36,405	31,052	33,761
9	-	1.20	-1.43	Correct	100,033	19,247	103,286	18,311
				Incorrect	20,005	10,715	19,103	9,300
10	-	1.92	-0.90	Correct	95,419	16,942	99,470	16,070
				Incorrect	17,974	19,665	17,305	17,155
11	-	0.57	-0.90	Correct	59,550	32,914	61,920	32,445
				Incorrect	33,283	24,253	32,685	22,950
12	***a*** **,** ***b***	0.48	0.01	Correct	40,411	36,644	41,011	36,548
				Incorrect	36,843	36,102	36,930	35,511
13	-	1.38	0.82	Correct	21,620	23,836	24,318	24,441
				Incorrect	23,303	81,241	24,067	77,174
14	-	1.11	-1.20	Correct	89,291	22,845	93,463	21,285
				Incorrect	23,569	14,295	22,309	12,943
15	***b***	0.28	0.09	Correct	41,406	36,724	42,203	36,887
				Incorrect	36,873	34,997	36,528	34,382
16	-	1.36	1.22	Correct	13,294	20,143	14,784	21,387
				Incorrect	19,385	97,178	20,224	93,605
17	-	0.25	0.35	Correct	35,280	36,762	36,061	36,938
				Incorrect	36,946	41,012	36,981	40,020
18	-	0.82	-0.18	Correct	47,609	32,516	50,236	32,320
				Incorrect	32,629	37,246	32,434	35,010
19	-	1.33	0.04	Correct	45,976	27,513	49,552	27,435
				Incorrect	27,963	48,548	27,675	45,338
20	-	1.27	-0.08	Correct	50,483	27,764	53,661	28,029
				Incorrect	28,056	43,697	28,050	40,260

(table continued in the next page)

Appendix 7 (continued)

Task	Parameter(s) changing over time	True parameters at Time 1		Original Response	No Changes in Ability Distribution		Ability Distribution Changing	
		a	*b*		Reference Response		Reference Response	
					Correct	Incorrect	Correct	Incorrect
21	-	0.96	1.17	Correct	16,029	26,036	17,862	26,720
				Incorrect	25,920	82,015	26,232	79,186
22	-	0.63	-0.88	Correct	61,475	32,428	63,644	31,763
				Incorrect	32,425	23,672	32,093	22,500
23	***a***	1.25	1.85	Correct	6,785	18,453	7,852	19,228
				Incorrect	17,622	107,140	18,748	104,062
24	-	1.52	-0.44	Correct	67,964	24,376	72,131	23,901
				Incorrect	25,067	32,593	24,393	29,575
25	-	1.06	-1.19	Correct	86,218	24,074	90,276	22,943
				Incorrect	24,729	14,979	23,654	13,127
26	***b***	0.63	-1.09	Correct	76,465	28,975	79,264	27,944
				Incorrect	29,192	15,368	28,269	14,523
27	-	0.87	0.08	Correct	40,721	32,251	43,706	32,006
				Incorrect	31,853	45,175	31,862	42,426
28	***a , b***	1.95	-0.97	Correct	109,472	14,663	112,235	13,656
				Incorrect	15,500	10,365	14,529	9,580
29	-	0.48	0.53	Correct	30,854	35,266	32,522	35,282
				Incorrect	34,865	49,015	35,201	46,995
30	-	2.61	0.94	Correct	19,089	13,610	21,944	14,270
				Incorrect	12,218	105,083	12,939	100,847

Note: *a* = Discrimination parameter
b = Difficulty parameter

Appendix 8

Number of correct and incorrect responses in the original vs. reference data, and results of the McNemar test for intervals along the continuum of estimated ability under both conditions (with and without changes in the test taker ability distribution over time)

a. Task 1 – ***incorrectly identified as drifting*** by the overall McNemar test

Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	*p*		Correct	Incorrect	χ^2	*p*
-3.0	Correct	35	0	0	3.00	.083	48	0	0	6.00	.014
	Incorrect		3	32				6	42		
-2.5	Correct	842	4	14	62.06	<.001	847	4	15	71.26	<.001
	Incorrect		97	727				109	719		
-2.0	Correct	3,595	79	322	148.32	<.001	3,655	127	328	183.74	<.001
	Incorrect		714	2,480				779	2,421		
-1.5	Correct	9,385	994	1,663	142.37	<.001	9,396	1,195	1,757	117.19	<.001
	Incorrect		2,426	4,302				2,460	3,984		
-1.0	Correct	17,878	4,737	4,030	52.16	<.001	17,685	5,132	4,016	68.54	<.001
	Incorrect		4,705	4,406				4,793	3,744		
-0.5	Correct	27,276	12,987	5,524	19.74	<.001	27,099	13,719	5,467	3.81	.051
	Incorrect		6,001	2,764				5,673	2,240		
0.0	Correct	31,692	21,666	4,708	17.48	<.001	31,875	22,690	4,394	19.76	<.001
	Incorrect		4,311	1,007				3,987	804		
0.5	Correct	27,261	22,495	2,468	36.76	<.001	26,932	22,763	2,237	63.71	<.001
	Incorrect		2,060	238				1,734	198		
1.0	Correct	17,699	16,093	865	17.84	<.001	18,070	16,628	801	25.57	<.001
	Incorrect		698	43				611	30		
1.5	Correct	9,887	9,459	249	12.54	<.001	10,090	9,716	238	30.37	<.001
	Incorrect		176	3				132	4		
2.0	Correct	3,862	3,788	53	13.84	<.001	3,799	3,744	43	17.47	<.001
	Incorrect		21	0				12	0		
2.5	Correct	588	583	5	5.00	.025	504	496	8	8.00	.005
	Incorrect		0	0				0	0		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	92,885	19,901	41.74	<.001	150,000	96,214	19,304	24.80	<.001
	Incorrect		21,212	16,002				20,296	14,186		

Note. True task parameters: $a = 1.44$, $b = -1.05$; highlighted row corresponds to the range of abilities which contains the difficulty parameter.
$df = 1$ for all McNemar tests

(continued in the next page)

Appendix 8 (continued)

b. Task 3 – ***correctly identified as non-drifting*** by the overall McNemar test

Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	*p*		Correct	Incorrect	χ^2	*p*
-3.0	Correct	35	0	0	2.00	.157	48	0	0	5.00	.025
	Incorrect		2	33				5	43		
-2.5	Correct	842	1	23	19.32	<.001	847	5	23	26.04	<.001
	Incorrect		64	754				73	746		
-2.0	Correct	3,595	30	211	56.38	<.001	3,655	28	237	38.63	<.001
	Incorrect		396	2,958				393	2,997		
-1.5	Correct	9,385	272	1,109	75.59	<.001	9,396	304	1,128	72.11	<.001
	Incorrect		1,558	6,446				1,569	6,395		
-1.0	Correct	17,878	1,305	3,153	58.09	<.001	17,685	1,373	3,274	52.31	<.001
	Incorrect		3,788	9,632				3,886	9,152		
-0.5	Correct	27,276	4,325	6,074	48.83	<.001	27,099	4,788	6,209	51.64	<.001
	Incorrect		6,869	10,008				7,036	9,066		
0.0	Correct	31,692	9,381	7,817	0.46	.495	31,875	10,488	7,768	1.25	.263
	Incorrect		7,732	6,762				7,629	5,990		
0.5	Correct	27,261	12,815	6,175	44.07	<.001	26,932	13,439	5,965	43.63	<.001
	Incorrect		5,459	2,812				5,265	2,263		
1.0	Correct	17,699	11,187	3,103	41.83	<.001	18,070	11,819	3,033	37.07	<.001
	Incorrect		2,614	795				2,577	641		
1.5	Correct	9,887	7,486	1,271	39.57	<.001	10,090	8,013	1,107	34.31	<.001
	Incorrect		973	157				848	122		
2.0	Correct	3,862	3,360	326	57.50	<.001	3,799	3,370	285	60.59	<.001
	Incorrect		159	17				127	17		
2.5	Correct	588	546	42	42.00	<.001	504	482	22	22.00	<.001
	Incorrect		0	0				0	0		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	50,708	29,304	1.62	0.203	150,000	54,109	29,051	2.17	0.141
	Incorrect		29,614	40,374				29,408	37,432		

Note. True task parameters: $a = 1.13$, $b = -0.14$; highlighted row corresponds to the range of abilities which contains the difficulty parameter.
$df = 1$ for all McNemar tests

(continued in the next page)

Appendix 8 (continued)

c. Task 5 – ***incorrectly identified as non-drifting*** by the overall McNemar test (true change in discrimination)

Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	*p*		Correct	Incorrect	χ^2	*p*
-3.0	Correct	35	0	0	6.00	.014	48	0	0	9.00	.003
	Incorrect		6	29				9	39		
-2.5	Correct	842	33	84	58.08	<.001	847	43	87	50.94	<.001
	Incorrect		216	509				210	507		
-2.0	Correct	3,595	312	585	72.21	<.001	3,655	401	623	62.12	<.001
	Incorrect		914	1,784				934	1,697		
-1.5	Correct	9,385	1,601	2,077	27.50	<.001	9,396	1,738	2,071	35.90	<.001
	Incorrect		2,429	3,278				2,475	3,112		
-1.0	Correct	17,878	4,482	4,317	11.09	.001	17,685	4,753	4,334	4.65	.031
	Incorrect		4,632	4,447				4,537	4,061		
-0.5	Correct	27,276	9,630	6,505	0.96	.328	27,099	10,178	6,361	0.89	.345
	Incorrect		6,617	4,524				6,468	4,092		
0.0	Correct	31,692	15,018	6,755	0.01	.925	31,875	15,617	6,727	0.84	.359
	Incorrect		6,766	3,153				6,621	2,910		
0.5	Correct	27,261	15,950	5,173	34.81	<.001	26,932	16,185	4,849	13.64	<.001
	Incorrect		4,590	1,548				4,492	1,406		
1.0	Correct	17,699	11,987	2,728	15.37	<.001	18,070	12,485	2,585	4.98	.026
	Incorrect		2,446	538				2,427	573		
1.5	Correct	9,887	7,518	1,181	6.98	.008	10,090	7,766	1,138	3.96	.047
	Incorrect		1,056	132				1,045	140		
2.0	Correct	3,862	3,251	335	14.30	<.001	3,799	3,196	341	16.08	<.001
	Incorrect		244	32				244	18		
2.5	Correct	588	535	46	30.77	<.001	504	475	29	29.00	<.001
	Incorrect		6	1				0	0		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	70,317	29,786	0.30	.581	150,000	72,838	29,145	1.70	.192
	Incorrect		29,922	19,975				29,462	18,555		

Note. True task parameters at Time 1: $a = 0.96$, $b = -1.04$; highlighted row corresponds to the range of abilities which contains the difficulty parameter at Time 1.
$df = 1$ for all McNemar tests

(continued in the next page)

Appendix 8 (continued)

d. Task 12 – ***incorrectly identified as non-drifting*** by the overall McNemar test (true changes in difficulty and discrimination)

Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	p		Correct	Incorrect	χ^2	p
-3.0	Correct	35	0	0	11.00	.001	48	1	8	3.85	.050
	Incorrect		11	24				18	21		
-2.5	Correct	842	77	142	13.84	<.001	847	77	156	11.52	.001
	Incorrect		212	411				222	392		
-2.0	Correct	3,595	459	773	11.95	.001	3,655	497	792	8.72	.003
	Incorrect		915	1,448				914	1,452		
-1.5	Correct	9,385	1,607	2,137	5.92	.015	9,396	1,619	2,218	4.05	.044
	Incorrect		2,299	3,342				2,354	3,205		
-1.0	Correct	17,878	3,460	4,416	1.03	.310	17,685	3,548	4,207	8.83	.003
	Incorrect		4,512	5,490				4,484	5,446		
-0.5	Correct	27,276	6,290	6,856	1.15	.284	27,099	6,417	6,758	0.10	.751
	Incorrect		6,731	7,399				6,795	7,129		
0.0	Correct	31,692	8,355	7,900	0.68	.410	31,875	8,657	7,927	0.23	.629
	Incorrect		8,004	7,433				7,988	7,303		
0.5	Correct	27,261	8,287	6,667	0.08	.782	26,932	8,292	6,685	2.92	.088
	Incorrect		6,699	5,608				6,489	5,466		
1.0	Correct	17,699	6,053	4,332	0.34	.561	18,070	6,147	4,402	0.14	.710
	Incorrect		4,278	3,036				4,437	3,084		
1.5	Correct	9,887	3,804	2,372	2.71	.100	10,090	3,855	2,413	1.22	.270
	Incorrect		2,260	1,451				2,337	1,485		
2.0	Correct	3,862	1,738	901	4.73	.030	3,799	1,655	851	1.51	.219
	Incorrect		811	412				801	492		
2.5	Correct	588	281	148	5.29	.022	504	246	131	7.21	.007
	Incorrect		111	48				91	36		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	40,411	36,644	0.53	0.465	150,000	41,011	36,548	1.98	0.160
	Incorrect		36,843	36,102				36,930	35,511		

Note. True task parameters at Time 1: $a = 0.48$, $b = 0.01$; highlighted row corresponds to the range of abilities which contains the difficulty parameter at Time 1.
$df = 1$ for all McNemar tests

(continued in the next page)

Appendix 8 (continued)

e. Task 15 – ***incorrectly identified as non-drifting*** by the overall McNemar test (true change in difficulty)

Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	p		Correct	Incorrect	χ^2	p
-3.0	Correct	35	0	0	12.00	.001	48	1	7	2.33	.127
	Incorrect		12	23				14	26		
-2.5	Correct	842	69	176	2.06	.151	847	100	174	7.00	.008
	Incorrect		204	393				227	346		
-2.0	Correct	3,595	501	745	23.67	<.001	3,655	541	781	3.90	.048
	Incorrect		945	1,404				861	1,472		
-1.5	Correct	9,385	1,617	2,205	2.68	.102	9,396	1,691	2,190	7.03	.008
	Incorrect		2,315	3,248				2,369	3,146		
-1.0	Correct	17,878	3,610	4,406	1.46	.228	17,685	3,631	4,380	0.01	.923
	Incorrect		4,520	5,342				4,389	5,285		
-0.5	Correct	27,276	6,425	6,932	1.62	.203	27,099	6,410	6,749	3.04	.081
	Incorrect		6,783	7,136				6,953	6,987		
0.0	Correct	31,692	8,681	7,957	1.22	.268	31,875	8,810	8,072	4.32	.038
	Incorrect		7,818	7,236				7,810	7,183		
0.5	Correct	27,261	8,383	6,795	0.52	.470	26,932	8,519	6,649	0.57	.449
	Incorrect		6,711	5,372				6,562	5,202		
1.0	Correct	17,699	6,187	4,277	0.50	.481	18,070	6,411	4,422	2.42	.120
	Incorrect		4,212	3,023				4,277	2,960		
1.5	Correct	9,887	3,899	2,328	0.61	.435	10,090	4,048	2,442	7.86	.005
	Incorrect		2,275	1,385				2,250	1,350		
2.0	Correct	3,862	1,741	907	3.50	.061	3,799	1,761	922	22.07	<.001
	Incorrect		829	385				731	385		
2.5	Correct	588	293	145	8.27	.004	504	280	99	1.07	.302
	Incorrect		100	50				85	40		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	41,406	36,724	0.30	0.585	150,000	42,203	36,887	1.75	0.186
	Incorrect		36,873	34,997				36,528	34,382		

Note. True task parameters at Time 1: $a = 0.28$, $b = 0.09$; highlighted row corresponds to the range of abilities which contains the difficulty parameter at Time 1.
$df = 1$ for all McNemar tests

(continued in the next page)

Appendix 8 (continued)

f. Task 23 – ***correctly identified as drifting*** by the overall McNemar test (true change in discrimination)

Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	*p*		Correct	Incorrect	χ^2	*p*
-3.0	Correct	35	0	0	-	-	48	0	0	1.00	.317
	Incorrect		0	35				1	47		
-2.5	Correct	842	0	4	1.33	.248	847	0	2	3.60	.058
	Incorrect		8	830				8	837		
-2.0	Correct	3,595	0	30	11.13	.001	3,655	1	44	4.77	.029
	Incorrect		62	3,503				67	3,543		
-1.5	Correct	9,385	8	215	12.55	<.001	9,396	8	224	16.62	<.001
	Incorrect		295	8,867				319	8,845		
-1.0	Correct	17,878	42	689	23.94	<.001	17,685	44	798	16.92	<.001
	Incorrect		883	16,264				971	15,872		
-0.5	Correct	27,276	163	1,852	14.02	<.001	27,099	208	1,990	18.66	<.001
	Incorrect		2,087	23,174				2,272	22,629		
0.0	Correct	31,692	488	3,412	7.50	.006	31,875	608	3,702	6.87	.009
	Incorrect		3,642	24,150				3,931	23,634		
0.5	Correct	27,261	1,072	4,363	0.25	.614	26,932	1,331	4,537	0.39	.534
	Incorrect		4,316	17,510				4,478	16,586		
1.0	Correct	17,699	1,563	3,813	11.46	.001	18,070	1,881	3,980	2.06	.151
	Incorrect		3,523	8,800				3,853	8,356		
1.5	Correct	9,887	1,771	2,682	66.52	<.001	10,090	1,941	2,753	56.68	<.001
	Incorrect		2,117	3,317				2,222	3,174		
2.0	Correct	3,862	1,307	1,176	127.17	<.001	3,799	1,498	1,136	147.62	<.001
	Incorrect		689	690				626	539		
2.5	Correct	588	371	217	217.00	<.001	504	332	172	172.00	<.001
	Incorrect		0	0				0	0		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	6,785	18,453	19.10	<.001	150,000	7,852	19,338	9.11	0.003
	Incorrect		17,622	107,140				18,748	104,062		

Note. True task parameters at Time 1: $a = 1.25$, $b = 1.85$; highlighted row corresponds to the range of abilities which contains the difficulty parameter at Time 1.
df = 1 for all McNemar tests

(continued in the next page)

Appendix 8 (continued)

g. Task 26 – ***incorrectly identified as non-drifting*** by the overall McNemar test (true change in difficulty)

Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	p		Correct	Incorrect	χ^2	p
-3.0	Correct	35	0	0	10.00	.002	48	0	0	13.00	<.001
	Incorrect		10	25				13	35		
-2.5	Correct	842	76	115	42.86	<.001	847	63	127	34.27	<.001
	Incorrect		238	413				239	418		
-2.0	Correct	3,595	586	766	22.88	<.001	3,655	605	772	23.37	<.001
	Incorrect		965	1,278				974	1,304		
-1.5	Correct	9,385	2,225	2,173	23.56	<.001	9,396	2,395	2,086	37.72	<.001
	Incorrect		2,505	2,482				2,502	2,413		
-1.0	Correct	17,878	5,888	4,202	8.28	.004	17,685	5,969	4,158	13.90	<.001
	Incorrect		4,470	3,318				4,505	3,053		
-0.5	Correct	27,276	11,468	6,213	0.11	.740	27,099	11,853	5,999	0.37	.542
	Incorrect		6,250	3,345				6,066	3,181		
0.0	Correct	31,692	16,346	6,439	0.01	.937	31,875	17,165	6,293	2.67	.102
	Incorrect		6,448	2,459				6,111	2,306		
0.5	Correct	27,261	16,668	4,812	12.90	<.001	26,932	16,844	4,556	4.81	.028
	Incorrect		4,466	1,315				4,349	1,183		
1.0	Correct	17,699	12,031	2,632	3.83	.050	18,070	12,815	2,453	2.66	.103
	Incorrect		2,492	544				2,340	462		
1.5	Correct	9,887	7,453	1,188	4.95	.026	10,090	7,865	1,111	9.00	.003
	Incorrect		1,082	164				974	140		
2.0	Correct	3,862	3,213	383	32.31	<.001	3,799	3,237	350	51.60	<.001
	Incorrect		241	25				184	28		
2.5	Correct	588	511	52	9.47	.002	504	453	39	14.29	<.001
	Incorrect		25	0				12	0		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	76,465	28,975	0.80	0.370	150,000	79,264	27,944	1.87	0.172
	Incorrect		29,192	15,368				28,269	14,523		

Note. True task parameters at Time 1: $a = 0.63$, $b = -1.09$; highlighted row corresponds to the range of abilities which contains the difficulty parameter at Time 1.
$df = 1$ for all McNemar tests

(continued in the next page)

Appendix 8 (continued)

h. Task 28 – ***correctly identified as drifting*** by the overall McNemar test (true changes in difficulty and discrimination)

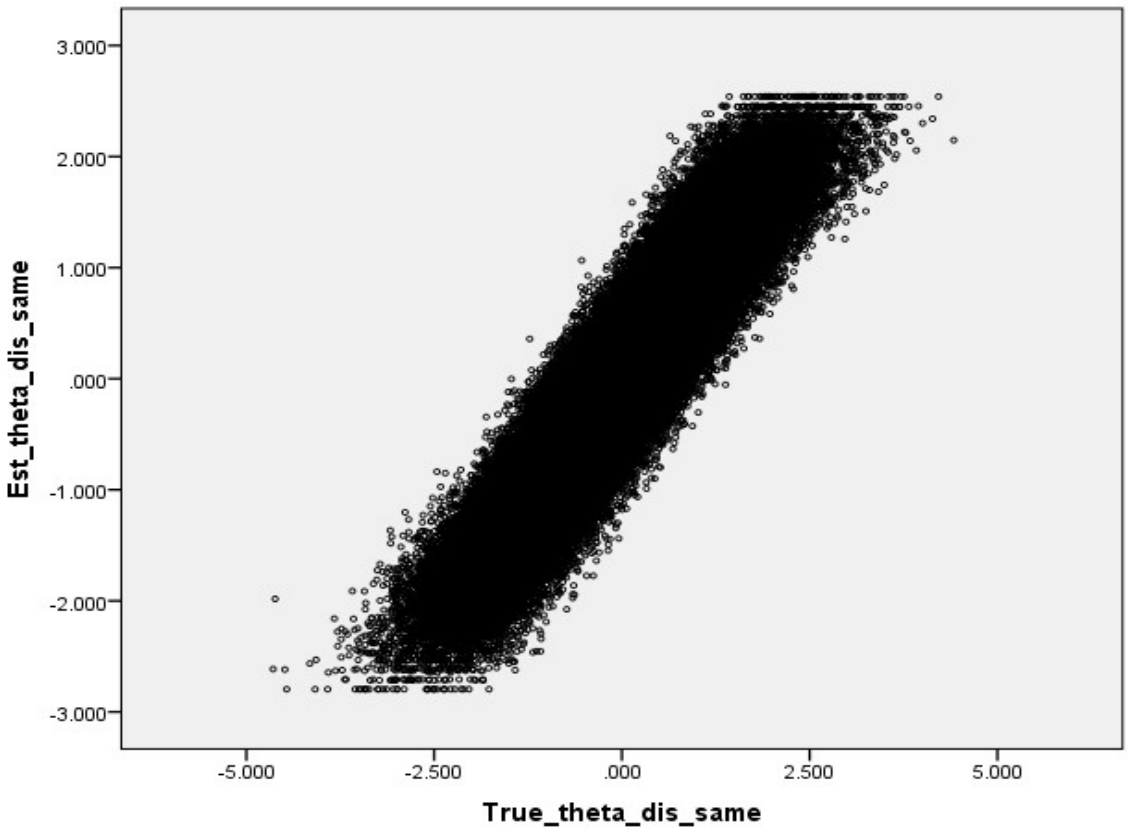
Estimated Ability (grouped)	Original Response	No Changes in Ability Distribution					Ability Distribution Changing				
		N	Reference Response		McNemar Test		N	Reference Response		McNemar Test	
			Correct	Incorrect	χ^2	*p*		Correct	Incorrect	χ^2	*p*
-3.0	Correct	35	0	0	3.00	.083	48	0	0	2.00	.157
	Incorrect		3	32				2	46		
-2.5	Correct	842	0	4	102.54	<.001	847	1	4	96.57	<.001
	Incorrect		114	724				108	734		
-2.0	Correct	3,595	113	343	187.21	<.001	3,655	126	354	193.94	<.001
	Incorrect		807	2,332				834	2,341		
-1.5	Correct	9,385	1,529	1,824	151.16	<.001	9,396	1,660	1,817	175.80	<.001
	Incorrect		2,646	3,386				2,709	3,213		
-1.0	Correct	17,878	6,999	3,905	35.42	<.001	17,685	7,521	3,719	31.37	<.001
	Incorrect		4,449	2,525				4,218	2,227		
-0.5	Correct	27,276	17,843	4,231	0.76	.382	27,099	18,531	3,945	2.95	.086
	Incorrect		4,151	1,051				3,794	829		
0.0	Correct	31,692	26,428	2,731	42.11	<.001	31,875	27,244	2,509	66.66	<.001
	Incorrect		2,272	261				1,963	159		
0.5	Correct	27,261	25,232	1,180	72.93	<.001	26,932	25,281	944	43.98	<.001
	Incorrect		800	49				677	30		
1.0	Correct	17,699	17,136	353	39.25	<.001	18,070	17,605	273	14.49	<.001
	Incorrect		205	5				191	1		
1.5	Correct	9,887	9,756	80	6.42	.011	10,090	9,982	74	14.81	<.001
	Incorrect		51	0				34	0		
2.0	Correct	3,862	3,848	12	7.14	.008	3,799	3,781	16	10.89	.001
	Incorrect		2	0				2	0		
2.5	Correct	588	588	0	-	-	504	503	1	1.00	.317
	Incorrect		0	0				0	0		
3.0	Correct	0	0	0	-	-	0	0	0	-	-
	Incorrect		0	0				0	0		
Overall	Correct	150,000	109,472	14,663	23.17	<.001	150,000	112,235	13,656	26.98	<.001
	Incorrect		15,500	10,365				14,529	9,580		

Note. True task parameters at Time 1: $a = 1.95$, $b = -0.97$; highlighted row corresponds to the range of abilities which contains the difficulty parameter at Time 1.
df = 1 for all McNemar tests

Appendix 9

Scatter plot of true test taker abilities vs. the abilities estimated using the proposed procedure for evaluating IPD (N=150,000)

a. No changes in the test taker ability distribution over time



(continued in the next page)

Appendix 9 (continued)

b. Test taker ability distribution changing over time