A Modern Assessment Psychometric Approach to Dynamic Assessment

Marié de Beer
University of South Africa

Address correspondence to Marié de Beer, Department of Industrial and Organizational Psychology, University of South Africa, P.O. Box 392, Pretoria 0003, South Africa. E-mail: dbeerm@unisa.ac.za

Although dynamic assessment (DA) has been hailed as a positive move towards fair assessment, it has generally not been used in educational or industry settings to the same extent that standard (static) tests have been. The present article attempts to elucidate how the use of Item Response Theory (IRT) and Computerised Adaptive Testing (CAT) can address some of the problems typically associated with dynamic assessment. An example of a DA tool that makes use of IRT and CAT, shows acceptable psychometric properties and is comparable to standard tests in terms of ease of administration illustrates the possibility of wider application of DA in both educational and industry settings.

Keywords: dynamic assessment (DA), item-response theory (IRT), computerised adaptive testing (CAT)

In the world history of psychometrics—and in particular in the domain of cognitive assessment—dynamic assessment (DA) has a very special place. DA refers to an assessment approach that includes a learning opportunity during assessment in order to provide information on the current as well as the potential future performance levels of the individual being assessed—typically by means of a test-train-retest process and with the aim of measuring both the level and rate of learning (Caffrey, Fuchs & Fuchs, 2008; Grigorenko, 2009; Lidz, 2009). Vygotsky (1978) is often referred to as the father of DA with his concept of the zone of proximal development (ZPD), but the work of Binet and Simon (1905; 1915) at the beginning of the 20th century can actually be regarded as the first empirical use of dynamic assessment principles for the measurement of learning potential. Binet and Simon were tasked by the French government of the time to develop a measure for conducting assessments among poorly performing school children, to distinguish those who could benefit from further training and educational interventions—to such an extent that they would be able to join normal educational settings—from those who would probably not benefit and would need to remain in special educational settings (Wolf, 1973).

Brief Overview

Despite its original dynamic test characteristics, Binet and Simon’s test came to be used as a static cognitive ability test and is often indicated as the forerunner of standard static cognitive tests as they are generally used today (Fancher, 1985; Van der Linden, 2008a; Wolf, 1973). A further interesting feature of the Binet-Simon test is that, in principle, it was also an adaptive test, with features such as variable entry level, continuous scoring and a termination rule based on performance (and accuracy of measurement) (Van der Linden, 2008a). In the light of research and work in the field of cognitive assessment that followed in the next 100-plus years, the groundbreaking contributions of these scientists can be appreciated even more.

Although DA has generated considerable interest and research, it has not been absorbed into mainstream assessment commensurate with expectations (Grigorenko & Sternberg, 1998; Lidz, 2009). Various authors have reviewed the successes, contributions and limitations of dynamic assessment (Beckmann, 2006; Caffrey et al., 2008; Grigorenko, 2009; Grigorenko & Sternberg, 1998; Lidz, 2009; Murphy & Maree, 2006; Sternberg & Grigorenko, 2002), Sternberg and Grigorenko (2002, p. viii) described dynamic assessment as “a wonderful idea whose implementation for the most part has been less than fully satisfactory”. Limiting factors noted by them are the dearth of published research on psychometric information about dynamic assessment tools, as well as practical elements such as measurement problems when comparing pre- and post-test scores and time-consuming administration. According to a number of authors (Grigorenko & Sternberg, 1998; Lidz, 2009), these factors have limited the wider implementation and use of dynamic assessment alongside standard static assessments in mainstream education and other domains. The aim of this article is to elucidate and demonstrate how the use of IRT and CAT can address a number of these limitations.

According to Sternberg and Grigorenko (2002, p. 30), “multiple attempts to quantify learning potential and to transform the construct of dynamic testing into a set of robust psychological diagnostic tools have not produced consistent results. Nevertheless, the idea of dynamic testing is so appealing that, despite its relatively sparse empirical validation, it has been widely discussed and fairly widely used.” They question whether this approach can be supported by results and techniques that show through empirical results that it can provide information “over and above the data collected by conventional tests” (Sternberg & Grigorenko, 2002, p. 181). In a meta analysis of available predictive validity results for DA measures, Caffrey et al. (2008) reported significant unique predictive validity shown by DA measures over and above that shown by traditional cognitive tests in predicting future academic achievement. However, Lidz (2009, p. 16) report that these methods are not generally used due to “insufficient supportive evidence and heavy time requirement.”

Murphy and Maree (2006) provided an extensive overview of dynamic assessment research done in South Africa. Other authors (Caffrey et al., 2008; Grigorenko & Sternberg; Sternberg & Grigorenko; 2002) have also reviewed work done in other regions in the world. It therefore does seem that there are selected areas or specific fields of application where DA is
used, but it has not yet seen the kind of general application that
is seen with standard (static) tests (Lidz, 2009). A partial expla-
nation for the limited application of DA could be that within the
field itself, different approaches and areas of application exist
which share certain commonalities, but which also differ in im-
portant ways.

**DA for Measurement of Learning Potential**

For the purpose of this article, two broad approaches to DA
are distinguished, namely a more clinical, diagnostic and reme-
dial approach, on the one hand, and a psychometric, measure-
ment-oriented, comparative approach, on the other. This dis-
tinction links to the ideographic versus the nomothetic
approaches (Hermans, 1988) and with specific reference to DA
was referred to as clinically-oriented DA versus research ori-
ented DA by Caffrey et al. (2008). With regard to DA, the two ap-
proaches share the following similarities: 1) both are based on
the Vygotskian concept of the Zone of Proximal Development
(ZPD); 2) both use a test-train-retest strategy—assuming that
standard tests do not reflect optimal performance levels; both
aim to identify learning potential; and 3) they both provide train-
ing relating to the assessment task. However, there are also
distinct differences between the two approaches. The first ap-
proach is more diagnostic, remedial, individually oriented and
typified by the Feuerstein model (Caffrey et al., 2008;
Grigorenko, 2009) and often involves assessment of low-per-
forming individuals. In this approach, procedures are not neces-
sarily standardised and the training or learning experience pro-
vided is determined by the ability profile, performance and
problem areas of the particular individual assessed. Since
inter-individual comparison is not the focus here, the lack of
standardisation in the learning experience does not represent a
limitation. The second approach is more measurement ori-
ented, comparative or norm-based and involves assessment of
individuals who perform within the normal range of ability. This
latter approach incorporates the principles of dynamic assess-
ment by providing a learning opportunity during assessment,
but it is more focused on accurate measurement, the
psychometric properties of the instruments used and ease of
administration to allow for group assessments, thereby making
these tools more comparable with standard assessment mea-
sures (Caffrey et al., 2008). Since the learning provided in this
approach is standardised, inter-individual comparison is possi-
bile. This more standardised, measurement-oriented dynamic
assessment approach provides the kind of empirical
psychometric evidence that addresses some of the concerns
noted by Grigorenko and Sternberg (1998).

**Standard or Static Tests**

Standard cognitive assessment has a long and extensive
history of use in a number of domains – for vocational guidance
in education, and in industry for screening and selection, for ap-
pointments and promotions as well as training and develop-
ment. These tests provide useful information for decision mak-
ing, as evidenced in the predictive validity values typically
found. However, in the last few decades there has been an in-
creased need for measures of learning potential to be used
alongside standard cognitive assessment (Lidz, 2009; Murphy & Maree, 2006). Standard (static) tests take into consideration
only current levels of performance and are based on an implicit
assumption that those who are tested had similar opportunities
to develop the constructs being measured. Among other things,
differences in socioeconomic and educational background have
an effect on the development of cognitive-related constructs in
particular (Claassen, 1997; Foxcroft, 1997; Paterson & Uys,
2005). Educationally and socioeconomically disadvantaged in-
dividuals are often at a further disadvantage when standard
cognitive tests are used, because these tests typically include
content representing crystallized abilities (ie language profi-
ciency, scholastic content or educational material) influenced
by prior learning experiences (Claassen, 1997; Foxcroft, 1997;
Van de Vijver, 1997, 2002). In South Africa in particular, differ-
ences in socioeconomic and educational levels (StatsSA, 2008)
contribute to social and political strife around employment and
educational and developmental opportunities.

By contrast, dynamic tests, with their focus on measuring
learning potential, make provision for a learning opportunity dur-
ing assessment, to allow for measurement of the projected fu-
ture or potential levels of performance (Caffrey et al., 2008;
Grigorenko, 2009). They can therefore provide information that
can be well used in decisions relating to screening and selection
as well as training and development opportunities. This ap-
proach takes into consideration the fact that many individuals
may have had less than optimal learning opportunities and
therefore might not yet have attained their optimal levels of per-
formance. The measurement of learning potential allows for
fairer assessment of disadvantaged individuals for two reasons:
firstly, the assessment includes a learning experience, and sec-
ondly, the test items typically measure fluid reasoning ability,
which is less influenced by socioeconomic and educational
background and prior learning experiences.

Research has indicated that intelligence quotient (IQ)
scores from static tests are influenced by educational opportu-
nity and socioeconomic factors and are subject to change
(Claassen, 1997; Grigorenko & Sternberg, 1998; Vincent,
1991). This has partly led to the increased attention that learn-
ing potential measures have received. Improved socioeco-
nomic and educational opportunities typically result in increases
in the mean IQ scores over and above normal population in-
creases over time (Van de Vijver, 1997; Vincent, 1991). Based
on a household survey conducted in 2007, with regard to educa-
tional attainment of persons aged 20 years and older, only
33.4% have completed a Grade 12 (senior secondary 23.6%) or
higher (9.8%) level of qualification (StatsSA, 2008). The high
proportion of this group with attained qualification levels lower
than a completed senior secondary level emphasizes the need
for the measurement of learning potential to provide training
and development opportunities on a fairer basis for individuals
at all educational levels.

**Learning Potential Assessment**

In order to utilize learning potential assessment for partici-
pants who perform within the normal range, Vygotsky's (1978)
work needs to be revisited. Vygotsky (1978) proposed dynamic
testing as a means of measuring what he referred to as the
ZPD, referring to the zone between an individual's independent
(actual) and guided or assisted (potential) levels of perfor-
mance. In Vygotsky's view, the task of assessment is to identify
not only those cognitive processes that are fully developed, but
also those that are in a state of being developed at the time of
assessment and which can be identified by incorporating learn-
ing as part of the assessment procedure (Dörfler, Golke, &
Artelt, 2009). The idea is that a dynamic test would measure a
person's ability to profit from learning opportunities, guidance
and assistance and thereby provide insight into his or her learn-
ing potential. The more the individual improves on his or her ini-
Implications for Practice

From a practical perspective, measurement of learning potential is typically used in the context of learning and training, with the logical result that actual developmental level (pre-test performance) also needs to be considered in learning potential assessment. Measurement of actual and potential levels of performance with consideration of the improvement in performance the individual shows following a learning experience represents a simulation. The assumption is that examinees are likely to show commensurate levels of performance and utilize learning opportunities in a similar fashion in real life.

The reason for assessment is important, since the question at hand is not merely whether individuals show improved performance following a training intervention. For practical purposes, what is important is whether individuals assessed currently (as evidenced in their pre-test level of performance) meet the level of reasoning skills required by the training or development opportunity, or whether they show that they have the potential (as indicated by their post-test level of performance) to reach the required level. The focus is not only on whether the individual will be able to profit from learning, but also on the level of training he or she will be able to cope with — or alternatively, to what degree he or she seems able to cope with a particular level of training offered. Thus in practical terms, current and projected future (potential) levels of performance are compared with the opportunity (i.e., target level of training) to evaluate whether the individual is currently already performing at the required level or shows the potential to perform at the required level.

Predictive validity for dynamic assessment across normal ability ranges uses the pre- and post-test scores as predictors respectively since the difference score is not expected to correlate with criterion performance measures, as explained by Jensen (1963, p. 1):

> when improvement with practice is thus measured from a different baseline for every subject, the results can be confusing and are often uninterpretable. A subject who is initially good at the task is already near the asymptote of his learning curve and can therefore show but little gain or improvement with practice. The slowest learners can often show the greatest gain. Consequently, correlations between gain scores on various learning tasks and psychometric measures of intelligence usually average close to zero.

Dynamic assessments are efficiently delivered using IRT approaches and with CAT administration. These applications are considered next.

The Use of IRT and CAT in DA

In IRT correspondence can be established between observed performance and individuals’ locations on the latent trait being measured (De Ayala, 2009). This permits accurate measurement of the latent trait being measured — and with regard to DA, also the accurate measurement of difference or improvement scores. Furthermore person and item parameters are theoretically invariant (De Ayala, 2009), meaning item properties are sample independent, while person ability parameter is item independent (Van der Linden, 2008a; Weiss, 1983). This, in combination with the key characteristic of IRT - that item difficulty and person ability are measured on the same scale — makes possible CAT, which involves ongoing interactive selection of items from a precalibrated item bank to continually match the respondent’s estimated ability level (Van der Linden, 2008a; Weiss, 1983). Frey and Seitz (2009) report that in CAT the number of items can be reduced between 50% and 70%, compared to a fixed number of items conventional test - without loss of measurement precision. The IRT-based CAT approach also counters the floor and ceiling effects often encountered in static testing.

IRT and CAT address a number of the concerns regarding dynamic testing (Embretson, 1996; Embretson & Reise, 2000; Sijtsma, 1993a; 1993b), with CAT allowing for shorter testing times and IRT providing improved accuracy of measurement of individual difference scores (Kim-Kang & Weiss, 2008), as well as improved means to compare the scores of the same or different examinees since observed performance is linked to location on the latent trait being measured. "Because of its ability to equate testings and link item pools onto a common metric, IRT has the potential of offering solutions to the problem of measuring gains in achievement levels during the process of instruction" (Weiss, 1980, p. 8). IRT and CAT procedures seem particularly appropriate for learning potential assessment because they improve both measurement accuracy and time efficiency resulting in testing times comparable to those of standard cognitive tests (Dörfler et al., 2008; Van der Linden, 2008b). This modern psychometric approach therefore addresses the further concerns noted by Grigorenko and Sternberg (1998) and Sternberg and Grigorenko (2002) as limiting the broader application of dynamic assessment and provides a modern-day solution to ensure fair, accurate and effective measurement of learning potential.

An Example Illustrating the Use of IRT and CAT in DA in the South African Context

The Learning Potential Computerised Adaptive Test (LPCAT) addresses some of the concerns that have been noted regarding dynamic assessment (De Beer, 2005). Three different types of nonverbal figurative item formats are used namely figure series, figure analogies and pattern completion. Since they measure fluid general reasoning ability, they do not rely on language proficiency or scholastic background and are deemed to be more equitable than items with verbal or scholastic content (Claassen, 1997; Foxcroft, 2004; Hugo & Claassen, 1991; Owen, 1998). The LPCAT uses a test-train-retest format with two separate but linked adaptive tests, in an attempt to measure learning potential in the fluid reasoning ability domain so that language proficiency or formal academic qualifications should not impact significantly on performance (De Beer, 2000a; 2000b). Factor analysis results indicate a definite unidimensional structure for the item bank with coefficient alpha values ranging between 0.926 and 0.978 for different groups.
Following IRT-based Differential Item Functioning (DIF) analyses, items that indicated DIF above a set level for sub-groups based on gender, level of education, first language and culture were discarded and not used in the final compilation of the LPCAT (De Beer, 2000b; 2005).

Assessment equity issues. In the South African context the Employment Equity Act legislation of 1998 requires empirical evidence of the reliability, validity and fairness of psychological assessment instruments. The LPCAT shows coefficient alpha reliabilities for the item bank ranging between 0.925 and 0.987. Construct and predictive validity values vary depending on the samples and specific comparison measures used but construct validity values ranging between 0.100 and 0.713 and predictive validity values ranging between 0.008 and 0.610 have been reported (De Beer, 2000b, 2006). Furthermore, the following features could be considered positive in terms of fairness in the multicultural and multilingual South African context:

- It is a power test that focuses on measurement of general non-verbal figural (fluid) ability and has no overall time limit.
- It is a dynamic test that allows for learning during assess-
- It makes use of IRT measurement principles which links per-
- It makes use of CAT, whereby items are administered to match the estimated level of performance throughout and which decreases testing time significantly.
- It requires the use of only the space bar and Enter key, for ease of administration and thus allows for assessment of indi-
- The test instructions are available in all 11 official South Afri-
- Efficient measurement with the LPCAT. The LPCAT incor-
- The LPCAT compares well with standard tests in terms of its psychometric properties, test administration time and ease of interpretation, and is considered fair for use with disadvantaged examinees (De Beer, 2000b; 2006). Construct and predictive validity results at various educational levels have been reported (De Beer, 2000b; 2006; Van der Merwe & De Beer, 2006).
Table 1

<table>
<thead>
<tr>
<th>LPCAT T-score range</th>
<th>LPCAT Stanine score</th>
<th>ABET* / NQF level</th>
<th>Educational</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 32</td>
<td>1</td>
<td>Abet level 1</td>
<td>Grade 0 – 3</td>
</tr>
<tr>
<td>33 – 37</td>
<td>2</td>
<td>Abet level 2</td>
<td>Grade 4 – 5</td>
</tr>
<tr>
<td>38 – 42</td>
<td>3</td>
<td>Abet level 3</td>
<td>Grade 6 – 7</td>
</tr>
<tr>
<td>43 – 47</td>
<td>4</td>
<td>Abet 4 / NQF 1</td>
<td>Grade 8 – 9</td>
</tr>
<tr>
<td>48 – 52</td>
<td>5</td>
<td>NQF level 1 – 3</td>
<td>Grade 10 – 12</td>
</tr>
<tr>
<td>53 – 54</td>
<td>6</td>
<td>NQF level 4 – 5</td>
<td>Grade 12+ (Higher Certificate)</td>
</tr>
<tr>
<td>55 – 57</td>
<td>6</td>
<td>NQF level 6</td>
<td>Diploma/Advanced Certificate</td>
</tr>
<tr>
<td>58 – 62</td>
<td>7</td>
<td>NQF level 7</td>
<td>3-year Degree/Adv. Diploma</td>
</tr>
<tr>
<td>63 – 68</td>
<td>8</td>
<td>NQF level 8</td>
<td>Honours / 4-year Degree/ Postgraduate Diploma</td>
</tr>
<tr>
<td>69 – 80 (65+)</td>
<td>9</td>
<td>NQF level 9</td>
<td>Advanced Degree (Master’s)</td>
</tr>
<tr>
<td>69 – 80 (65+)</td>
<td>9</td>
<td>NQF level 10</td>
<td>Advanced Degree (Doctorate)</td>
</tr>
</tbody>
</table>

*Abet: Adult Basic Education and Training

Reflective Summary

To return to the initial aim, this article attempted to elucidate how IRT and CAT can be used to address a number of problems associated with DA and explained how these modern assessment principles and methods were used in the construction of the LPCAT.

The LPCAT has shown satisfactory psychometric properties and is comparable with standard tests in terms of ease of administration and testing time, thereby also addressing those concerns mentioned by Grigorenko and Sternberg (1998). It can be administered individually or in groups, has standard instructions, makes use of IRT-based CAT which enhances measurement accuracy and testing time, and because it is computerised, the results are available immediately on completion of the assessment. The fact that the pre- and post-test scores can be directly linked to benchmarked educational levels improves the practical utility of the results.

It serves as a practical example of the implementation of IRT and CAT to enhance the utility of DA alongside standard tests in educational and industry settings.

References


