MASTERARBEIT | MASTER'S THESIS

Titel | Title

Revisiting Guthke's process-oriented concept of psychological assessment: A proof-of-concept study in longitudinal data

verfasst von | submitted by

Lisa Marie Kasal BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Betreut von | Supervisor:	Assoz. Prof. Mag. Dr. Jakob Pietschnig Privatdoz.

Theoretical background

Learning

In the context of cognitive psychology, learning is defined as a relatively enduring change in behaviour or behavioural potential resulting from experience. This definition entails three essential characteristics: (a) learning involves a genuine change, (b) this change is relatively stable over time, and (c) it occurs through experience rather than temporary influences such as fatigue or mood (Shuell, 1986, as cited in Schunk, 2012). While explicit instruction involves conscious and deliberate efforts to acquire new knowledge or skills, learning can also occur without such intentional guidance. This latter form, referred to as *implicit learning*, is typically characterized by two features: (a) it is an unconscious process and (b) it yields abstract knowledge. *Implicit knowledge* results from inducing an abstract representation of the structure present in the stimulus environment, acquired in the absence of conscious, reflective strategies to learn. *Explicit learning* involves the conscious and intentional acquisition of knowledge, typically accompanied by the learner's awareness of what has been learned and the ability to articulate the underlying rules or structures. In contrast to implicit learning, explicit learning relies on deliberate hypothesis testing, rule formation, and the use of working memory resources (Sun, Slusarz, & Terry, 2005).

Intelligence and its measurement

The term *Intelligence* has been defined in various ways: A definition by Wechsler (1944) describes intelligence as the overall capacity to act purposefully, think logically, and adapt effectively to one's environment. Comparably, Neisser et al. (1996) defined intelligence as the capacity to understand complex ideas, to reason, and to adjust to changing environmental demands and emphasized adaptive problem-solving and learning from experience. Beyond definitions, various theoretical models have been developed to explain how intelligence is structured, and which abilities form its core. Among the earliest was Spearman's (1904) g-factor theory, which proposed that a single underlying ability (g) contributes to performance across cognitive tasks. In contrast, Thurstone (1938) suggested that intelligence consists of several independent "primary mental abilities", such as verbal comprehension, numerical reasoning, and spatial visualization, which he regarded as relatively independent from one another. Cattell (1963) introduced the distinction between *fluid intelligence*—the capacity for abstract reasoning and solving novel problems independent of prior knowledge—and *crystallized intelligence*, which reflects accumulated vocabulary, factual knowledge as well as culturally transmitted skills (Cattell, 1963).

One of the central insights of later research was that both perspectives capture important aspects of cognition: specific abilities exist, however they are intercorrelated and can

be explained by higher-order factors (e.g., Carroll, 1993; Schneider & McGrew, 2012). This synthesis is reflected in hierarchical models of intelligence: Carroll's (1993) three-stratum theory describes intelligence as a layered system with a general factor at the top, broad abilities (such as fluid and crystallized intelligence) at the second level, and more specific cognitive processes at the third. The most influential contemporary framework, the Cattell–Horn–Carroll (CHC) model, integrates these traditions and organizes intelligence into a comprehensive structure that includes several broad abilities—such as fluid reasoning (Gf), crystallized knowledge (Gc), visual-spatial processing (Gv), short-term memory (Gsm), and processing speed (Gs)—alongside more specific skills (Schneider & McGrew, 2012).

Standardized intelligence tests aim to operationalize theoretical models in applied settings. Figural matrix measures, such as Raven's Progressive Matrices, are frequently applied to capture aspects of fluid reasoning (Pallentin et al., 2024), whereas vocabulary-based instruments such as the Mehrfachwahl-Wortschatz-Intelligenztest (MWT-A; Lehrl, 2005) and the Wortschatztest (WST; Schmidt & Metzler, 1992) are considered indicators of comprehension-knowledge (Gc) according to the CHC model. Broader test batteries, including the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Wechsler, 2008) and the Kaufman Assessment Battery for Children – Second Edition (KABC-II; Kaufman & Kaufman, 2004), are designed to measure a wide range of broad and narrow abilities and are supported with hierarchical models such as the CHC framework (Flanagan & Harrison, 2012).

Status intelligence tests

Intelligence is conventionally measured using *status-oriented tests*, which refer to assessments in which the examiner presents tasks to the test taker and records their responses without making any effort to intervene, adjust, or enhance the performance of a test taker (Guthke & Wiedl, 2003). These tests are widely employed in educational, clinical, and organizational contexts, typically in the form of commercially available assessments used for diagnostic or placement decisions (and offer standardized and objective measurement conditions (Neisser et al., 1996). Status-oriented assessments typically provide a snapshot of current performance (*status diagnostics*) rather than evaluating an individual's capacity to learn or adapt. Furthermore, these tests often fail to provide useful guidance for remediation or targeted intervention. Especially in the educational context, there is a persistent communication gap between teachers and psychologists: Test results from status-oriented intelligence testing have no information on how to improve and teachers are unsure how to deal with such information and what value this information offers in terms of shaping the learning process (Tzuriel, 2000; Börnert-Ringleb, 2018). Such tests may underestimate the abilities of individuals whose

performance is constrained by factors such as limited educational opportunities or unfamiliarity with the testing format (Grigorenko & Sternberg, 1998). This diagnostic limitation has motivated the development of more dynamic approaches.

The concept of Dynamic Testing

Dynamic assessment (DA) and dynamic testing (DT) can be understood as components of a broader concept that encompasses various forms of assessment and testing. These approaches integrate feedback, prompts, or training into the testing process and are designed to capture a person's progress in solving cognitive tasks, in order to provide an indication of their potential for learning (Resing et al., 2020).

A distinction can be drawn between DA and DT: DA is primarily used in clinical settings and focuses on intervention. In this approach, the type and amount of feedback or mediation are determined individually, depending on the child's needs. Support is provided whenever a child is unable to solve a task independently (Elliott et al., 2018). In contrast, DT is mainly applied in research contexts and relies on standardized procedures for both test administration and feedback provision in order to compare participants in an adequate way. The main goal is to elicit intraindividual variability in performance, with the assumption that interindividual differences in this variability provide a more accurate reflection of the dynamics underlying human behavior, such as learning and cognitive flexibility (Guthke & Wiedl, 1996). Another important goal of DT is to identify and eliminate non-intellectual obstacles such as milieu- or socioeconomic-related disadvantages (Kubinger, 2019) that hinder the expression of one's own intelligence, thereby revealing abilities that are already present but often not recognized by traditional, status-oriented ability tests (Haywood & Tzuriel, 2002). Moreover, DT is not designed to replace traditional status testing, but to serve as an additional procedure. It can offer hypotheses for specific questions and deliver information that other measures cannot provide (Lidz, 1987), for example, it offers details on how individuals benefit from training (Tiekstra et al., 2016).

This concept is discussed in different forms in international literature (e.g. Veerbeek & Vogelaar, 2024, Wood et al., 2024) as an alternative to traditional status tests, especially in intelligence testing. The interest in DT is partly due to dissatisfaction with status-oriented assessment methods, for example they tend to underestimate abilities from minority populations (Resing et al., 2020). The principle of DT can be traced back to the work of Guthke and Wiedl (1996), who sought to overcome the limitations of status intelligence tests and other conventional psychometric approaches. The "classical learning test", a specific and the most recognized form of DT, involves a pretest, a training or instructional phase and a posttest (Guthke,

1992). In the pretest, baseline performance is established using tasks of a certain difficulty. During the training or instructional phase, the examiner provides feedback, prompts, or graduated hints (i.e. stepwise levels of assistance during problem solving) designed to simplify problem solving and promote understanding of task principles. Finally, in the posttest, participants complete new or parallel items under similar conditions to evaluate learning gains and the degree of modifiability that occurred during the intervention (Guthke & Wiedl, 1996). Another form of DT is "Testing-the-limits", in which examiners provide systematic prompts during the posttest, for example by asking metacognitive questions (e.g., "What goes through your mind as you try to solve this task?" or "Why do you consider this option more plausible than the others?") and by offering corrective feedback when errors occur (Wiedl, 2003). Guthke (1982) emphasized that a true understanding of intelligence requires moving beyond measuring the status quo to exploring the capacity for change and adaptation. A typical feature of this approach is the move away from the usual psychometric test, which only measures a momentary level of performance towards a diagnostic procedure in which changes are evoked and recorded in the test process in order to validate and make fairer statements about characteristics and abilities and their changeability (Guthke & Wiedl, 1996). This approach intends to provide incremental validity over status intelligence tests, particularly for individuals from disadvantaged backgrounds or those with irregular learning histories, as it provides a more dynamic individualized measure of cognitive potential (Guthke & Beckmann, 2003).

The theoretical foundations of DT were established much earlier. For example, Thorndike (1924) defined intelligence as the ability to learn and proposed that measuring this ability should be an essential component of assessing intelligence. This early conceptualization laid the groundwork for subsequent approaches that placed learning at the focus of intelligence assessment.

Zubin (1950) stated that psychological research should prioritize the analysis of individual behavioral patterns over group-level statistics: He proposed a set of axioms outlining this idiographic perspective. According to these principles, each person should be regarded as an independent universe that can only be compared or grouped with others after being examined in detail. Zubin further argued that every individual has a unique level of performance, and any observed test score merely represents a random sample of that individual's broader potential. Both individuals and their psychological characteristics exhibit inherent variability, the magnitude of which differs from person to person (Zubin, 1950).

Building on this perspective, DT is also rooted in the work of Luria (1976) and Vygotsky (1978). Both researchers emphasized the role of social and cultural factors in cognitive

development, arguing that the capacity to learn cannot be fully understood through status assessments alone. Luria's (1976) studies demonstrated how cognitive processes adapt according to social contexts, suggesting that intelligence is flexible and shaped by environmental demands. The author supported the idea that human consciousness is a product of social development, and it underlines that cognitive abilities are flexible and develop in response to social and environmental interactions. The author's method, which involved observing real-time learning and adaptive processes, shifted the focus from status-oriented measurement to understanding how individuals learn and respond to interventions, which paved the way for viewing intelligence as adaptable and later genuinely influenced diagnostic techniques (Luria, 1976).

Vygotsky (1978) further introduced the concept of the Zone of Proximal Development (ZPD), which highlights that a person's cognitive ability can only be fully understood when their capacity to learn with support is examined, rather than by status-oriented test performance alone.

The Theory of Structural Cognitive Modifiability (SCM), developed by Reuven Feuerstein (1980), is another key precursor to DT: It is based on the idea that cognitive structures are not static, but rather modifiable and developable. This theory postulates that cognitive abilities are not solely determined by innate factors or early experiences but can be influenced by targeted interventional processes that alter thinking patterns and learning processes in a sustainable way. Feuerstein argued that every individual — regardless of their current level of performance — has the capacity to improve their cognitive abilities through mediation and specific learning methods.

In DT, this broader capacity is captured by the concept of *learning potential* (LP), which symbolizes the ability to improve performance through structured training and practice. In general, LP is assessed using a test–train–test design, in which pre–post differences indicate the extent to which support fosters change, often expressed in numerical or categorical indices (Boosman et al., 2016). Thus, LP reflects the degree—and occasionally the speed—of improvement achieved under facilitation and serves as an indication of a person's modifiability.

As outlined above, one of the major criticisms of status intelligence testing is its limited ability to capture LP in individuals from culturally or socioeconomically diverse backgrounds. In such cases, status-oriented scores may underestimate genuine capabilities. For example, Stevenson et al. (2016) demonstrated in their study that DT testing can be particularly useful in revealing the cognitive potential of children from disadvantaged and ethnic minority groups—contexts in which traditional conventional measures often fail. Similarly, Zaaiman et

al. (2001) demonstrated the value of DT in selection procedures, especially when applicants originate from educationally disadvantaged environments. However, a unified concept of DT does not exist within the current body of research. Instead, it serves as an umbrella term for a range of different approaches (Börnert-Ringleb, 2018).

DT tests often incorporate inductive reasoning tasks (e.g., Resing, 2017; Touw et al., 2019). Inductive reasoning is widely regarded as a core component of fluid intelligence (Vo & Csapó, 2022). Tests assessing crystallized intelligence are less susceptible to learning effects (Guthke & Wiedl, 1996).

Within this framework, prominent instruments have been developed, including the Learning Potential Assessment Device (LPAD, Feuerstein et al.,1979), and the Adaptive Computerized Intelligence Learning Test (Guthke, 1995).

After Guthke's death in 2004, research activity in this field declined (Beckmann, 2004). Still, DT continues to offer a promising and equitable framework for assessing LP. Recent research has pushed the boundaries of DT in several promising directions. A notable example is the *Training-Only Dynamic Test* (ToDT) developed by Veerbeek and Vogelaar (2024), which dispenses with the traditional pretest and posttest phases and instead uses a standardized training (graduated prompts) session alone to infer a learner's instructional needs, suggesting that even minimal intervention designs may hold diagnostic value.

Computerized and adaptive formats are also gaining traction. For instance, Vogelaar et al. (2021) introduced a process-oriented computerized dynamic test for primary school mathematics which integrates real-time feedback and measuring time and planning dynamics.

Procedure in DT: The Learning Test Approach

The present study focuses on the Learning Test approach, developed by Guthke (1972) as the central procedure within DT. This approach functions as follows: Two assessments of the target competency are conducted. The first assessment follows a traditional diagnostic approach. The pretest collects information about the student's current performance level on the required task and can be regarded as a status-oriented test. The training phase in between pre- and posttest makes it possible to observe learning progress and individual differences in the learning process. After the pretest, participants receive targeted support, usually in the form of graduated prompts or feedback. The participant's updated level of performance is assessed during the post-test session. In the posttest, equivalent items are used to assess the extent to which participants have improved their performance as a result of the training. An improvement in post-test results is considered evidence of learning. Guthke et al. (2003) even questioned whether the value of the second measurement is not also a better description in terms

of status diagnostics than the initial value because unequal previous experiences could be compensated for by the support provided. Thus, it could deliver a more accurate representation of the actual target competence, reflecting a higher level of test fairness (Guthke, Beckmann & Wiedl, 2003).

Disadvantages of testing with instruction

Despite its strengths, DT has practical constraints that need to be considered. For example, DT requires more time to conduct and demands a higher level of skill, improved training, greater experience, and increased effort compared to status-oriented testing. For example, in the context of school exams, resources such as time and personnel are often limited, which may lead school psychologists to prioritize quick and effective solutions over more time-intensive methods like DT (Guthke et al., 2003; Haywood & Tzuriel, 2002).

Furthermore, DT has not become part of routine assessment practice and status-oriented testing remains the prevailing paradigm in psychological testing. Questions remain about the stability of difference scores and the use of external criteria such as school achievement or teacher ratings (Guthke & Wiedl, 1996).

The specific procedures of DT are mainly determined by the points at which the "instructional phase" is incorporated into the testing process and the methods used to evaluate the effects, which means that even straightforward feedback can be delivered in various ways (Guthke & Wiedl, 1996). This discussion about how much instruction should be provided within DT and the methodological complexities involving feedback naturally leads to a further question: Could comparable diagnostic insights—namely, indicators of an individual's LP—also emerge under simpler conditions, even without explicit feedback?

Repeated exposure to comparable problem types can already lead to measurable performance gains, suggesting that some aspects of LP may be observable through retesting alone (Raven, 1956; Guthke & Wiedl, 1996). Examining these effects provides a valuable starting point for understanding whether the fundamental principle underlying DT can also be captured through test repetition as a type of DT.

Cognitive Mechanisms

Performance gains can emerge from repeated testing even in the absence of explicit training. Several cognitive mechanisms are thought to underlie such improvements. One of the most robust findings in cognitive psychology is the *spacing effect*, which shows that learning opportunities distributed across time foster better long-term retention than massed practice (Cepeda et al., 2006). Repeated engagement with similar problem types can also enhance strategy use and processing efficiency, giving rise to *practice effects* (Hausknecht, Halpert, Di

Paolo, & Moriarty Gerrard, 2007; Salthouse & Tucker-Drob, 2008). Importantly, these gains often remain task-specific and show limited transfer to other cognitive domains, as highlighted by meta-analyses on working memory training (e.g. Ripp et al., 2022, Rodas et al., 2024). Related processes such as priming and processing fluency may additionally contribute by facilitating recognition and reducing cognitive load, thereby allowing faster and more accurate responses (Tulving & Schacter, 1990; Reber et al., 2004). *Retest effects*—systematic score increases when the same or an equivalent test is administered more than once—are typically attributed to increased familiarity with the test format, refined problem-solving strategies, and reduced cognitive effort (Hausknecht et al., 2007; Salthouse & Tucker-Drob, 2008).

While these mechanisms explain short-term improvements, they do not fully address whether such gains represent genuine cognitive modifiability. Although related, LP and practice effects are distinct phenomena. Practice effects are most pronounced in the initial retest trial and largely reflect task familiarity (Hausknecht et al., 2007), whereas LP is indexed by the slope of improvement across multiple learning opportunities (Guthke & Wiedl, 1996). Hammer et al. (2024) underline that practice effects are visible across trials but not slopes, with the first trial typically showing the strongest benefit at follow-up and ceiling effects constraining subsequent gains.

Psychometric Challenges of Equivalent Test Forms

A common strategy to control for practice effects is the use of equivalent forms (often also referred to as parallel forms) as performance gains are less attributable to item-specific familiarity and more to general task familiarity. However, constructing truly equivalent forms and ensuring their psychometric comparability is notoriously difficult, as it greatly increases demands on test development and validation (Schmidt et al., 2005). This makes it necessary to examine whether performance gains differ between identical and equivalent versions at the ability level (fluid and crystallized intelligence), as this distinction helps to determine whether improvements reflect item-specific practice or genuine learning.

The purpose of this study

The central aim of this study is to determine whether repeated testing – using either identical or comparable test forms - can be considered an independent form of learning intervention and therefore as a method of DT. Traditionally, DT incorporates structured feedback and instructional support to actively facilitate learning. However, Guthke and Wiedl (1996) noted that while it is often claimed, it is rarely empirically demonstrated that learning occurs during test engagement. The crucial diagnostic idea underlying DT is that individuals differ in the degree to which they profit from instruction or feedback (Guthke & Wiedl, 1996). These

interindividual differences in “test learning” are regarded as indicators of cognitive modifiability and, consequently, as a reflection of a person’s LP rather than their static performance level. The effects of this procedure without intervention have yet to be thoroughly examined. Additionally, practice effects represent a significant challenge for the interpretation of repeated testing, and equivalent test forms are often considered a potential remedy. However, as outlined above, truly equivalent forms are difficult to construct and may differ in terms of item difficulty, reliability, or validity. To address this gap, a proof-of-concept study was conducted employing a within-subjects design with repeated measures (two testing time points) to investigate LP. In particular, this study examines whether improvements occur when participants are re-exposed to identical or equivalent test items after a delay, suggesting that they may have implicitly learned from their prior engagement with the tasks. Such between-session gains would indicate that elements of test learning—and thus LP —can emerge even without explicit instruction. In the present study, two testing conditions were implemented: an identical condition, in which participants completed the same test form at both measurement points, and a non-identical but comparable condition, in which they completed a different test measuring the same cognitive domain (crystallized intelligence) than in the first time point. The LP is expected to be stronger in the identical condition, as practice effects incorporate both item-specific familiarity and general task learning, whereas improvements in the non-identical condition are limited to the transfer of strategies and format familiarity.

Methodology

Prior to collecting any data, the study design, analysis plan, and specific primary hypotheses were pre-registered on the Open Science Framework (OSF, https://osf.io/7v2t5, registered and last updated on October 24th, 2024). In addition, the study was approved by the Departmental Review Board of the Faculty of Psychology at the University of Vienna.

Hypotheses

Based on the research questions, the following hypotheses were formulated:

1. Respondents show a significant improvement in their test performance between the first and second test time point.

1a. Respondents show a significant improvement in their test performance between the first and second test time point if the second test is an identical repetition of the first test.

1b. Respondents show a significant improvement in their test performance between the first and second test time points if the second test is a non-identical measure assessing the same cognitive domain as the first test.

2. The observed learning potential, operationalized as performance improvement from Time 1 to Time 2, is greater when the identical test is repeated than when a comparable, non-identical version is administered.

Research Design and Procedure

Participants

An a priori power analysis ($\alpha = .05$, power = .80 expected effect size $d = 0.20$) conducted with the R package "pwr" (Champely, 2020) indicated a minimum required sample size of $n = 258$ participants for within-subjects comparisons. Because the study required participants to complete two testing sessions two weeks apart, a higher attrition rate was expected. Therefore, the target sample size was increased to $n = 300$ to compensate for potential dropouts between sessions. In total, 183 participants were recruited. Participants were required to be enrolled in the psychology bachelor's program at the University of Vienna to partake in this study to ensure a homogeneous educational background. This approach helped to keep participants' educational backgrounds comparable, reducing potential confounding influences. Students were recruited through The Laboratory Administration System for Behavioral Science (LABS) that is used to manage participation in experiments for psychology students from the University of Vienna. They were granted credits after finishing the testing which they can use to complete a mandatory course (Faculty of Psychology, 2024). Participants were given the option to end the testing early or to opt out of the assessment points. Of the initial

183 participants, 20 did not attend the second session and as stated in our preregistration, were therefore excluded from the analyses, which resulted in a final sample of 163 valid cases (65.03% women; age: $M = 22.51$, $SD = 4.44$, range: 18 – 43 years).

The study employs a within-subjects design with repeated measures. In this design, each participant experienced two systematically varied conditions which allows to assess the effects within the same individuals. This approach removes between-subject variability and therefore enhances the sensitivity of our findings.

At the first measurement point, all participants were exposed to the same initial condition. Two weeks later, at the second measurement point, the conditions were systematically varied. Specifically, the presentation order of the second condition was counterbalanced to control for order effects. Half of the participants experienced Condition A first and Condition B second, while the other half experienced Condition B first and Condition A second. Condition A was designed so that participants first completed the same test items as they had two weeks earlier, followed by equivalent items. Condition B had the reverse order: Participants first completed the equivalent version, then the same items as the first measurement.

Materials used

To assess both fluid and crystallized intelligence, four standardized tests were administered: the *Standard Progressive Matrices* (SPM; Raven, 2009), the *Advanced Progressive Matrices* (APM; Raven, 1998), the *Wortschatztest* (WST; Schmidt & Metzler, 1992)and *the Mehrfachwahl-Wortschatztest, Version A* (MWT-A; Lehrl et al., 1974). The SPM assess individuals' cognitive abilities across a broad spectrum and are largely independent of demographic factors such as age or education, thus testing fluid intelligence. The SPM consist of five sets of twelve items each, progressively increasing in difficulty. Test subjects fill in the missing parts of the representations by choosing the appropriate option from the six available answers (Raven, 1938). More precisely, individuals are required to infer the underlying rules that determine how elements and attributes are organized across the matrix, and to apply these rules to select the correct completion —a process that involves *correspondence finding* and *rule indication.* This test is widely used for research purposes and in clinical educational settings (Carpenter et al., 1990) as well as in developmental research (Yip et al., 2025) which highlights its widespread global application, particularly due to its nonverbal and culture-reduced design. Furthermore, the SPM have demonstrated adequate reliability, with Cronbach's alpha coefficients ranging from .84 to .90 across age groups (Sbaibi et al., 2014). In the present sample, internal consistency for the SPM was $\alpha = .47$ at T1, $\alpha = .56$ for the identical version at T2, and $\alpha = .44$ for the equivalent version at T2 (95% CIs [.33, .58], [.45, .65], and

[.30, .56], respectively). These values fall below the conventional threshold for acceptable internal consistency (George & Mallery, 2003). However, as Ellis (2013) notes, optimal reliability depends on the experimental design and may be lower for short group-based tasks.

One major disadvantage of the SPM is the occurrence of ceiling effects, where higher- and lower-performing individuals may not be adequately discriminated due to the test's limited difficulty (Pind et al., 2003). To address this issue, the APM were included in the procedure as well. This test is typically administered to adults with higher ability levels and consists of two sets in which set II has a greater difficulty level than set I (Carpenter et al., 1990). Set I consist of 12 test items whereas set II consist of 36 items (Raven, 1938). Identical to the SPM, the tasks involve incomplete geometric figures or patterns that must be completed by selecting one of eight alternative answers. This test has demonstrated solid psychometric properties, with the findings from Nurhudaya et al. (2019) using Rasch analysis on 12,343 Indonesian students reporting high internal consistency (Cronbach's α = .85), person reliability of .86, and item reliability of 1.00. This supports both the reliability and unidimensionality of the instrument.

The WST was chosen to assess crystallized intelligence, specifically verbal knowledge. This test is a standardized vocabulary test, with Rasch-scaled ability values and corresponding IQ norms available. The WST consists of 42 test items, arranged according to increasing difficulty. It is used to briefly assess verbal intelligence level and language comprehension. Participants must identify the real word from a set of options, which includes one correct word and several pseudowords (non-existent words). One item consists of five distractors and one target term and there is no time limit. The WST demonstrates high reliability in validating studies, with a split-half coefficient (Spearman–Brown) of $r = .95$ and an internal consistency (Cronbach's α) of .94 (Schmidt & Metzler, 1992). In the present sample, the internal consistency of the WST items was moderate (Cronbach's α = .68) (George & Mallery, 2003).

The MWT-A was chosen as a conceptually comparable measure to the WST. The MWT-A follows a similar structure and procedure as the WST because both tests use the same format of selecting the correctly spelled real word among pseudowords and it has 37 items. Evidence for validity comes from strong correlations with established intelligence measures, such as the WST, with reported coefficients of up to .95 (Lehrl et al.,1971). It uses four distractors and one target word in each item (Lehrl et al., 1991). Administration and scoring are fully standardized and can be conducted without extensive professional training. The test has also demonstrated high reliability, with split-half coefficients typically ranging

between .89 and .95 and a test–retest reliability of *r_tt* = .95 (Lehrl et al., 1974). In the present sample, the internal consistency of the test, estimated using the split-half method with Spearman–Brown correction, was *r* = .64. According to the conventional benchmarks proposed by George and Mallery (2003), this value indicates questionable internal consistency.

During a single session, participants completed 9 items from the SPM, 8 items from the APM, and all 42 items from the WST. As mentioned earlier, at the second measurement point, participants receive a comparable, non-identical version of the initial items. For the SPM, 9 items were taken from the five difficulty sets (A–E). Each set represents a distinct level of complexity, with items increasing in difficulty from Set A to Set E. The shortened version spanned a wide array of difficulty levels by selecting items from each set. For the construction of the equivalent versions of the SPM and APM, the same test framework was used, but different items were drawn from the same original item sets so that the identical and equivalent items can be comparable in structure and content.

As part of a collaborative research project, additional intelligence mindsets were assessed using implicit and explicit test items on the computer at the end of the testing procedure. These measures, however, fall outside the focus of this thesis and were not included in the analysis.

Procedure

All tests were administered on a computer, using the *Qualtrics* survey platform (https://www.qualtrics.com/de/) (Qualtrics, Provo, UT). The first data collection point took place from November 18th to November 23rd, 2024, the second data collection from December 2nd to December 7th, 2024. Due to an insufficient number of participants in the first round, a second data collection period was carried out. The first data collection of the second round took place between January 20th, 2025, and January 22nd, 2025, and the second data collection was carried out between February 3rd, 2025, and February 5th, 2025. On both second measurement points, the participants were randomly assigned to either of the two test versions (identical or non-identical test items first) assembled by Qualtrics. All participants received the same set of items in both test versions. The randomization procedure referred exclusively to the order of item presentation which guaranteed that each participant worked on all items while controlling for sequence effects. The sequence of test presentation was identical across participants at T1 (SPM first, followed by APM and ending with WST). At T2, participants were randomly assigned to two groups differing in test sequence: Participants assigned to condition A completed the identical versions of the tests (SPM, APM, WST) first, followed by the comparable versions (SPM, APM, MWT-A). Participants assigned to

condition B completed the comparable versions first, followed by the identical versions. Each session took approximately one hour to complete, with a maximum of 20 participants being tested at the same time. No time limit was imposed in the sessions. The data analysis was performed with R version 2023.12.1+402 (R Core Team, 2023). A significance threshold of $p < .05$ was established for all analyses.

Data Transformation

Raw data from both measurement points were first matched using the unique linking question provided by each participant so that responses from the first (T1) and second (T2) testing sessions can be combined into a single record per participant. Participants were allowed to skip items; unanswered items were scored as incorrect (i.e., no credit was given).

For each cognitive test – the SPM, the APM in both identical and non-identical versions and the identical WST – sum scores were calculated separately for T1 and T2. Difference scores were then computed for each participant by subtracting the T1 score from the corresponding T2 score ($\Delta = T2 - T1$). These difference scores served as the dependent variables for the statistical analyses and were calculated in raw points, as the maximum achievable score was identical across measurement points.

For the direct comparison between the MWT-A (37 items) and the WST (42 items), scores were converted to percentages to account for differences in the maximum possible score between the two instruments.

To test whether participants improved from T1 to T2 overall (H1), under identical repetition (H1a), and under a non-identical same-domain measure (H1b), six paired sample *t*-tests were conducted separately for the identical and non-identical test versions using the scores from T1 and T2 to examine whether performance significantly improved from T1 to T2. For Hypothesis 2, the difference scores of the identical and non-identical versions were directly compared using paired-samples *t*-tests.

Prior to each *t*-test, the normality assumption for the difference scores was evaluated using the Shapiro–Wilk test. In all cases, the assumption was violated; therefore, the non-parametric Wilcoxon signed-rank test with continuity correction was additionally performed. Given the large sample size and the robustness of the t-test to deviations from normality, paired-samples *t*-tests were reported as the primary analyses, with Wilcoxon signed-rank tests included as robustness checks. Since the Wilcoxon test tests the median, not the mean, the results are not identical to the *t*-tests, but they showed a consistent pattern. Effect sizes for paired designs were calculated as Cohen's *d*, defined as the mean of the difference scores divided by their standard deviation. Effect sizes of 0.2, 0.5, and 0.8 were interpreted as small,

medium, and large (Cohen, 1988). Additionally, to each Cohen's *d* value, a confidence interval was calculated. All statistical analyses were conducted with a significance threshold of $\alpha = .05$ (two-tailed).

Results

Table 1 provides an overview of the descriptive statistics (means and standard deviations) for all cognitive measures at T1 and T2. These descriptive statistics provide an overview of participants' average performance levels. Mean scores on the identical versions of the SPM, APM, and WST increased from T1 to T2, whereas the comparable versions yielded little or no gains. For the MWT-A and the comparable APM, performance even declined.

Table 1

Descriptive Statistics for Cognitive Measures at T1 and T2, Including Mean Differences (ΔM) and Standard Deviations of Differences (ΔSD)

Test	*M*	*SD*	Δ*M* (T2–T1)	ΔSD
APM T1	6.22	1.41		
APM T2 (equivalent)	6.08	1.34	-0.14	1.50
APM T2 (identical)	6.50	1.42	0.28	1.44
SPM T1	7.04	1.53		
SPM T2 (equivalent)	7.21	1.29	0.16	1.58
SPM T2 (identical)	7.78	1.35	0.74	1.18
WST T1	30.86	3.50		
WST (identical)	31.2	3.6	0.8	4.94
MWT-A T2 (equivalent)	24.55	3.33	-7.12%	7.69%

Note. M = mean, SD = standard deviation, ΔM = mean difference, ΔSD = standard deviation of differences.

Paired-samples t-tests

For the APM equivalent version, the comparison between T1 and T2 indicated no significant difference in performance, $t(162) = -0.85$, $p = .398$, $d = -0.07$, 95% confidence interval (CI) [-0.22, 0.09], which signifies a trivial effect. The assumption of normality of the difference scores was tested using the Shapiro–Wilk test and revealed a violation. Therefore, the Wilcoxon signed-rank test was conducted, and it yielded a consistent result, $V = 2989$, $p =$

.320. This finding suggests that solving the equivalent version of the APM did not lead to systematic performance improvements across the sample.

In contrast, when participants repeated the identical APM version at T2, their scores were significantly higher compared to T1, $t(162) = 2.57$, $p = .011$, $d = 0.20$, CI [0.05, 0.36]. The size of this effect can be interpreted as trivial to small, indicating that practice with the exact same items provided a measurable, though modest, performance benefit. Due to the violation of the normality assumption of the difference scores, the results were additionally performed with the Wilcoxon signed-rank test ($V = 3924$, $p = .01$).

A similar pattern emerged for the SPM. The comparison between the equivalent SPM versions at T1 and T2 did not yield a statistically significant difference, $t(162) = 1.34$, $p = .182$, $d = 0.11$, CI [-0.05, 0.26]. As the Shapiro-Wilk-test also revealed a violation of the normality assumption, the Wilcoxon signed-rank test was conducted and was consistent with the *t*-test results, $V = 3828.5$, $p = 0.2199$, pointing to no performance changes when different but structurally similar items were presented.

However, when the identical SPM version was repeated at T2, participants showed a marked improvement, $t = 8.0979$, $df = 162$, $p < .001$, $d = 0.63$, CI [0.46, 0.80]. The effect size was in the medium range, demonstrating that repeated exposure to identical items significantly enhanced performance. The Wilcoxon signed-rank test yielded a consistent result, $V = 5446.5$, $p < .001$. As seen on Figure 1, participants demonstrated the largest improvement when repeating the identical SPM, whereas performance gains in the equivalent APM condition were negligible.

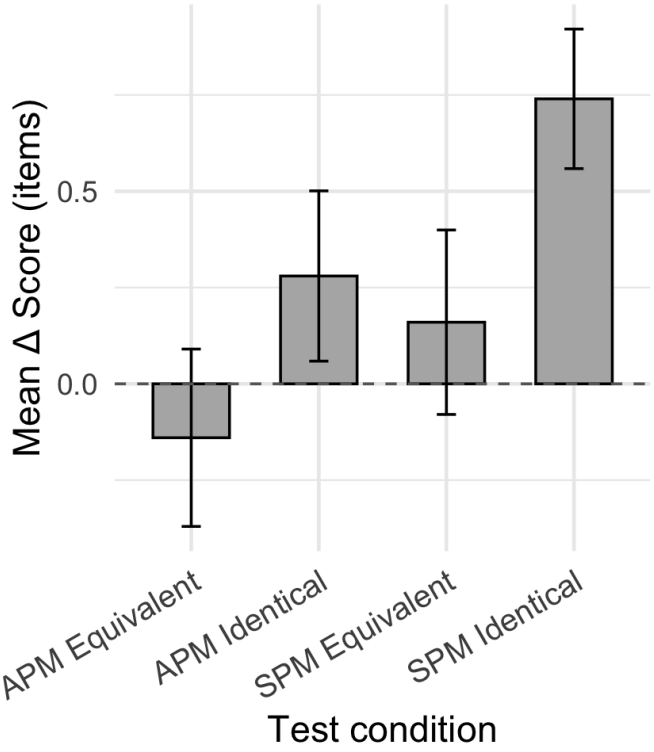

Figure 1. Mean performance changes (Δ = T2 – T1, in items) for identical and equivalent versions of the SPM and the APM. Bars represent the 95% CI of the mean.

Concerning the measurement of the verbal ability, the WST showed a small but statistically significant gain upon repetition: Scores at T2 were higher than at T1 when the identical WST version was administered, $t(162) = 2.08$, $p = .039$, $d = 0.16$, 95% CI [0.01, 0.32], which signifies a trivial to small effect. As the Shapiro-Wilk-test showed a violation of the normality assumption of the difference scores (T2-T1), the results of the t-test align with the Wilcoxon signed-rank test, $V = 4056.5$, $p = 0.02$. This suggests that even for a vocabulary recognition task, significant improvement occurs with repeated exposure. In contrast, the MWT-A yielded significantly lower scores than the WST at T2, $t(162) = -11.78$, $p < .001$, $d = -0.93$, 95% CI [-1.11, -0.74]. In this case, the normality assumption was also violated. The Wilcoxon signed-rank test yielded a consistent result, $V = 1128$, $p < .001$. This large negative effect indicates that the MWT-A has a higher difficulty than the WST. Figure 2 shows that the repetition of the WST led to a slight increase in performance, whereas with the MWT-A the performance decreased significantly.

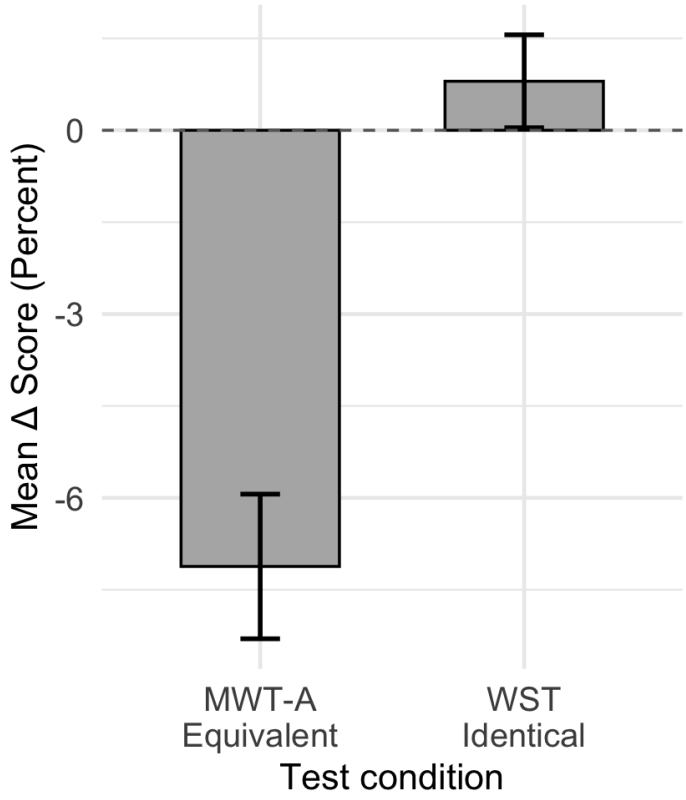

Figure 2. Mean performance changes (Δ = T2 – T1, in percent) for the WST and the MWT-A. Bars represent the 95% CI of the mean.

Paired samples-t-tests

Finally, direct contrasts between the identical and equivalent versions confirmed these differences. For the SPM, performance gains were significantly larger when participants repeated the identical version compared to the equivalent version, $t(162) = 5.9311$, $p = <.001$, $d = 0.41$, 95% CI [0.25, 0.57], representing a small to medium effect. Due to the violation of the normality assumption, a Wilcoxon signed-rank test was conducted, which supported the result, $V = 5279.5$, $p < .001$.

An analogous finding emerged for the APM, where the improvement was greater in the identical than in the equivalent condition, $t(162) = 4.06$, $p < .001$, $d = 0.32$, 95% CI [0.16, 0.47], corresponding to a small effect. The Wilcoxon test yielded a consistent result, $V = 4479.5$, $p = < .001$.

A paired-samples *t*-test comparing the difference scores of the WST (identical version) and the MWT-A (non-identical version) indicated that participants gained on the WST, but declined on the MWT-A. This difference in change was significant, $t(162) = 13.47$, $p < .001$ The effect size can be interpreted as large, $d = 1.06$, 95% CI [0.86, 1.25]. This finding was similar to the result of the Wilcoxon signed-rank test, which also indicated a significant

difference, $V = 12417$, $p < .001$. All these results suggest that the learning effects seen are mainly linked to the repetition of items, while equivalent versions produce little or no systematic enhancement. As Figure 3 illustrates, the performance gains in the identical conditions were systematically higher than in the equivalent versions.

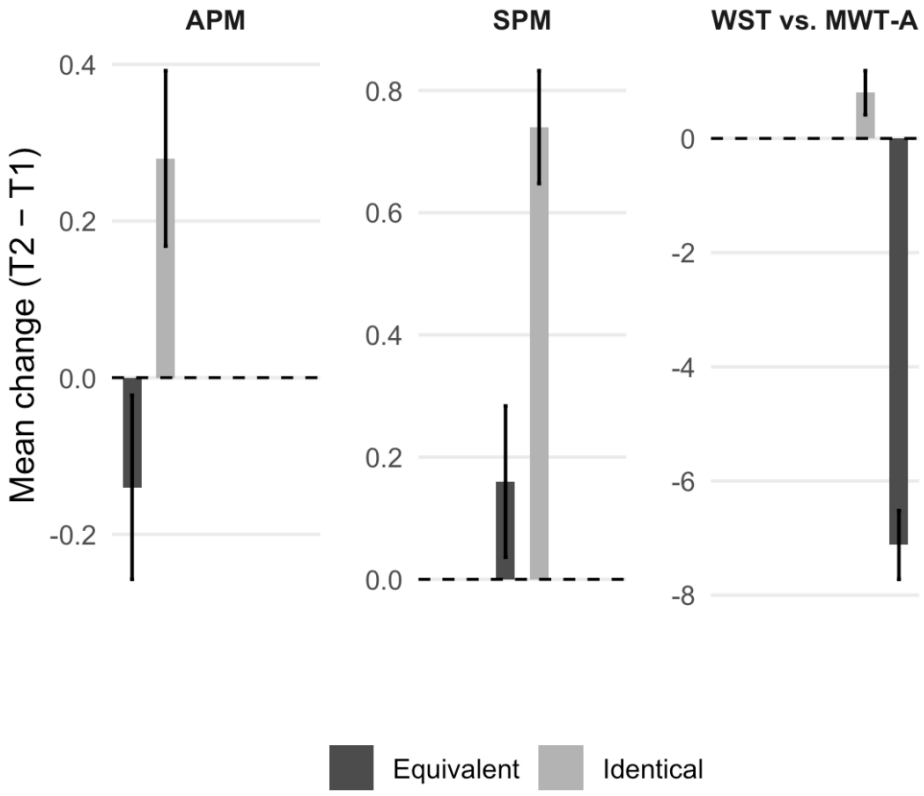

Figure 3. Direct contrasts of performance changes (Δ = T2 − T1) between identical and equivalent versions. For SPM and APM, changes are shown in number of items; for WST vs. MWT-A, changes are shown in percent. Bars represent the 95% CI of the mean. For APM and SPM, changes are expressed in number of items; for WST and MWT-A, changes are expressed in percent.

Additional Analysis

As the additional ANCOVA analysis showed, baseline crystallized ability (MWT-A) slightly moderated the retest gains, though not significantly. The analyses comparing the scores from WST T1 and MWT-A T2 showed that, despite targeting the same construct of crystallized intelligence, performance on the MWT-A was lower than on the WST, suggesting that the MWT-A is comparatively more difficult. Therefore, a repeated-measures ANCOVA was conducted with Time (T1, T2) as a within-subjects factor and z-standardised MWT-A scores as a covariate.

There was a significant main effect of Time, $F(1, 160) = 4.39$, $p = .038$, $\eta_p^2 = .03$, indicating a small overall improvement in WST performance from T1 ($M = 73.5$, 95% CI [72.5, 74.5]) to T2 ($M = 74.3$, 95% CI [73.3, 75.3]).

The main effect of MWT-A entered as a time-invariant covariate representing participants' general verbal ability, was highly significant, $F(1, 160) = 130.11$, $p < .001$, $\eta_p^2 = .45$, showing that individuals with higher crystallized ability performed better overall across both time points.

The critical Time × MWT-A interaction was not statistically significant, $F(1, 160) = 3.15$, $p = .078$, $\eta_p^2 = .02$, indicating that the degree of improvement did not reliably vary as a function of crystallized ability.

Although this interaction did not reach conventional levels of significance, the effect size was small but non-trivial. Considering that the study was underpowered for detecting interaction effects, this pattern can be interpreted as indicative. It suggests that participants with higher crystallized ability may have shown slightly greater gains from test repetition, an effect that could reach significance in a larger sample.

Exploratory follow-up analyses using estimated marginal means revealed that the WST gain was negligible for participants with low MWT-A scores (−1 SD: $\Delta = 0.12$, $p = .82$), modest but significant for those with average MWT-A scores ($\Delta = 0.81$, $p = .038$), and largest for those with high MWT-A scores (+1 SD: $\Delta = 1.50$, $p = .007$).

Overall, the pattern implies that participants with higher crystallized ability tended to benefit slightly more from test repetition which is consistent with the notion that prior knowledge may facilitate recall and familiarity effects.

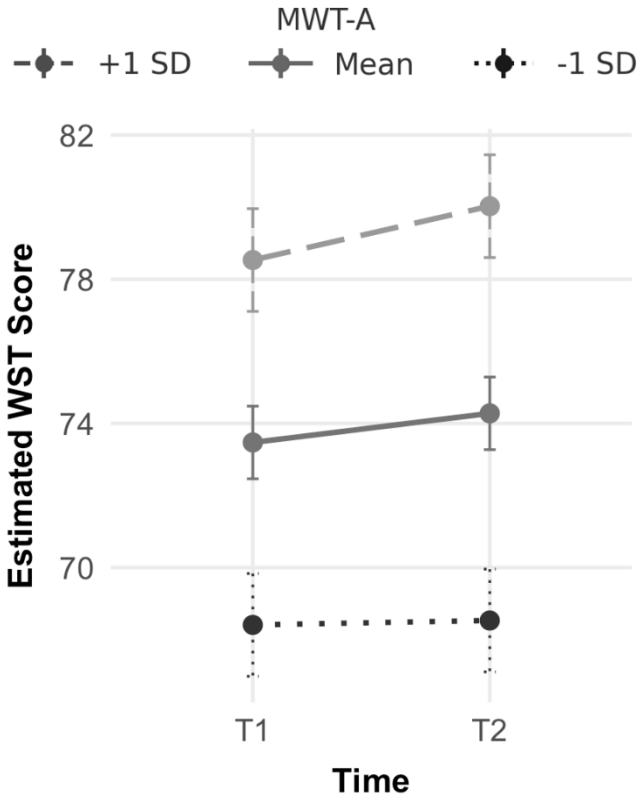

Figure 4. Estimated WST means (±95% CI) at T1 and T2 for participants with low (−1 SD), mean (0), and high (+1 SD) MWT-A scores.

Exploratory Analysis

As per the preregistered exploratory descriptives, participants were required to report their experience with the test administered, if applicable, in order to investigate any correlation between prior experience with the administered tests and variations in LP. A total of n = 13 participants reported prior experience: WST (n = 8), APM (n = 2), SPM (n = 2), and MWT-A (n = 1). Given the small and uneven group sizes, no inferential statistical analyses were conducted. Instead, descriptive statistics are presented in Table 2. While participants with prior experience in the WST appeared to show slightly higher gains than those without experience, the numbers are too small to draw meaningful conclusions.

Table 2

Comparison of mean scores and standard deviations between participants with and without prior test experience.

Test	Group	M	SD
Identical SPM	Experience	2	0
Identical SPM	No experience	0.83	1.21
Equivalent SPM	Experience	2	0
Equivalent SPM	No Experience	0.14	1.58
Identical APM	Experience	-1	1.41
Identical APM	No Experience	0.3	1.44
Equivalent APM	Experience	-1.5	2.12
Equivalent APM	No experience	-0.13	1.49
WST	Experience	1	1.85
WST	No experience	0.3	2.09
MWT-A	Experience	-10%	0
MWT-A	No experience	-7%	7.71

Note. Scores for MWT-A are expressed as percentages to allow comparability between instruments with different maximum points.

As part of the preregistered exploratory analyses, it was also investigated whether participants' native language (German vs. other) was associated with differences in LP. As with prior test experience, this variable was included in the preregistration as a potential predictor, particularly for the verbal tests (WST, MWT-A). Only $n = 13$ participants reported a native language other than German. Given the very small and uneven group sizes, identical to the exploratory descriptives about experience vs. no experience with the administered tests, no inferential statistical analyses were conducted. Instead, descriptive statistics for the WST and MWT-A gains are presented in Table 3. Due to the very small group sizes, these results have no inferential value and are reported for completeness only.

Table 3

Comparison of mean scores and standard deviations between native German speaking and non-native German speaking participants.

Test	Group	M	SD
WST	Native German	0.34	2.09
WST	Non-native German	0.15	1.95
MWT-A	Native German	-6.91%	7.74
MWT-A	Non-native German	-9.53%	6.95

Discussion

The present study was motivated by the framework of DT, which combines pretest, intervention, and posttest phases to assess LP. Against this background, it was examined whether mere test repetition—without feedback, prompts, or instructional support—can already be regarded as a form of intervention. Overall, the findings revealed practice effects when participants were retested with the same version of a test, whereas no significant performance gains emerged when non-identical versions were administered.

Hypothesis 1a: Identical test repetition leads to performance improvement

As predicted, participants showed significant performance gains when retested with the identical version of the APM, SPM, and WST. These findings align with the notion of practice effects, whereby repeated exposure to the same tasks facilitates familiarity with item content, reduces cognitive load, and improves response efficiency (Hausknecht et al., 2007; Salthouse & Tucker-Drob, 2008). The effect sizes observed for Raven's Matrices ranged from trivial to small for the APM and reached a medium level for the SPM, suggesting that item-specific solution strategies were successfully recalled and reused. The verbal domain, namely the WST, also showed small but reliable practice effects, highlighting that repetition benefits extend beyond reasoning tasks to crystallized knowledge measures.

Hypothesis 1b: Equivalent test versions lead to performance improvement

Participants did not significantly improve when confronted with equivalent versions of the APM and SPM, which appears to contradict Guthke and Wiedl's (1996) observation that even without instructional support, repeated engagement with tasks can evoke measurable improvements. The absence of effects implies that practice benefits may be very specific to particular items and do not extend to new problems that are structurally similar. In line with this, previous research has shown that retest gains are strongest for identical forms and substantially smaller when equivalent forms are used (Hausknecht et al., 2007). Interestingly, performance on the MWT-A even declined at T2, which suggests a higher difficulty of this measure compared to the WST. This highlights the importance of test characteristics: Improvements cannot be assumed when tests within the same cognitive domain differ systematically in difficulty or design.

Hypothesis 2: Identical test repetition produces larger learning potential than equivalent tests

The direct comparisons supported the expected pattern, with significantly larger performance gains in the identical condition than in the equivalent condition. In both the APM and SPM, practice effects were significantly larger in the identical condition than in the

equivalent condition. However, these findings do not indicate genuine learning in the sense of DT as described by Guthke (1992) but rather reflect classic practice effects. The improvements likely stem from increased familiarity with item formats and strategies, rather than from enhanced cognitive modifiability. In DT terms, this indicates that repetition alone cannot be equated with a full-fledged LP assessment: While improvements are observed, they do not reflect generalizable cognitive modifiability as conceptualized by Guthke (1992). Rather, the observed practice effects seem to operate on a narrower level of processing fluency and recall, consistent with cognitive mechanisms such as reduced novelty (Tulving & Schacter, 1990; Reber et al., 2004).

The observed practice effects in this study are consistent with prior meta-analytic evidence on retesting in cognitive ability assessments. Hausknecht et al. (2007) demonstrated that performance gains are particularly pronounced when identical test forms are administered repeatedly, whereas improvements are considerably smaller with equivalent forms. Importantly, their analyses further revealed that the magnitude of practice effects decreases as the retest interval increases, suggesting that short time spans between assessments amplify retest gains. Similarly, recent evidence from memory-based learning tasks has shown that practice effects tend to be strongest during the initial trials of a retest session and may diminish across subsequent learning trials (Hammers et al., 2024). This suggests that repetition primarily facilitates early retrieval or recognition processes. The findings of Melby-Lervåg & Hulme (2013) further suggest that practice effects may be transient, as they observed that improvements in working memory were not sustained over time. Similarly, the results of the present study show that although test repetition leads to immediate performance gains, these effects do not necessarily reflect deeper cognitive modifiability that would generalize beyond the tested items.

In addition, the present study employed a short retest interval of two weeks, which likely contributed to the relatively strong improvements observed for identical test versions.

Additional analysis

The ANCOVA was used to explore whether individual differences influenced the size of the retest effects. Verbal ability, measured by the MWT-A, was included as a covariate to examine whether general crystallized intelligence explained additional variance in post-test performance beyond initial scores. The results showed a positive but non-significant trend, suggesting that participants with higher verbal ability tended to show slightly larger gains (ηp^2 = .03, small effect; Cohen, 1988). Although this effect did not reach statistical significance, the pattern aligns with prior research showing that practice effects can vary as a function of

initial ability (Hausknecht et al., 2007; Ackerman, 1988). Given the limited statistical power of the present design, this finding should be interpreted cautiously, as the observed trend may still reflect meaningful differences in learning or familiarity rather than pure measurement error.

Moreover, the trend-level interaction between baseline crystallized ability and retest gains ($p = .078$; partial $\eta^2 = .02$) points in a theoretically meaningful direction. Although this effect did not reach statistical significance, its non-trivial magnitude suggests that participants with higher verbal ability may have benefited slightly more from test repetition. This pattern aligns with the idea that existing knowledge structures can facilitate recall or strategic processing when identical material is presented again. While the present study was underpowered to detect such interaction effects conclusively, future research with larger samples could clarify whether baseline crystallized knowledge systematically moderates the size of practice-related gains.

Practical Implications

The findings of the present study have several practical implications for both research and assessment. First, they demonstrate that repeated administration of identical test versions can lead to performance gains. This points to the possibility of artificially elevated ability scores if the same test items are reused in diagnostic practice. In high-stakes settings such as personnel selection or clinical diagnostics, it is therefore advisable to rely on psychometrically equivalent test forms to minimize inflated scores caused by simple repetition effects.

At the same time, the study shows that tests testing the same cognitive domain are not automatically comparable. In this case, performance on the MWT-A was even lower compared to the WST, likely due to systematic differences in difficulty and construction. For practical implications, this underlines the need to evaluate the psychometric equivalence of equivalent forms rather than assuming they can be used interchangeably.

Moreover, the results indicate that practice effects are genuine but highly item specific. Improvements in identical conditions do not essentially reflect a transfer of learning or broader cognitive modifiability. For psychological research, this means that practice-related gains should not be misinterpreted as indicators of general LP. Instead, DT approaches that incorporate structured feedback and instructional support are needed to capture transferable learning processes more accurately.

In the context of DT, the present findings highlight an important conceptual distinction. Practice effects observed in repeated identical tests reflect improvements that are highly item-bound, arising from recognition, strategy recall, and reduced novelty (Tulving &

Schacter, 1990; Reber et al., 2004). In contrast, LP refers to the broader capacity to acquire new strategies or knowledge that generalize beyond the practiced material (Guthke & Wiedl, 1996; Tzuriel, 2001). The results of this study therefore confirm previous conclusions that repetition alone cannot be interpreted as evidence of genuine modifiability (Sternberg & Grigorenko, 2002; Hausknecht et al., 2007). Instead, it represents a narrower form of task-specific adaptation consistent with findings from retest research showing short-term gains without enduring cognitive change (Salthouse & Tucker-Drob, 2008; Melby-Lervåg & Hulme, 2013). Taken together, these results suggest that DT captures a qualitatively different form of learning—one that depends on guided feedback and transfer rather than on repeated exposure. Although the effects in the present study did not reach statistical significance, their overall pattern points in the direction proposed by Guthke and Wiedls (1996) framework: Learning and cognitive modifiability appear to unfold most clearly when facilitative support is provided. In this sense, the findings neither contradict Guthke's assumptions nor fully confirm them; instead, they indicate that test repetition alone can elicit limited performance gains, while the diagnostic strength of DT lies in its ability to reveal broader modifiability under conditions of structured feedback. Guthke and Wiedl (1996) emphasized that learning tests differ fundamentally from static measures by capturing intraindividual variability and modifiability under facilitation. By isolating simple test repetition, this study shows that improvements can occur even without instructional support, yet these gains represent mere retest effects rather than the kind of learning or cognitive modifiability described by Guthke and Wiedl (1996). In this sense, the findings provide empirical support for the notion that the added value of DT lies precisely in its structured instructional component, which enables transfer beyond item-specific familiarity. The results align with previous research showing that the diagnostic strength of DT depends on the presence of feedback and guided support: Earlier studies (e.g., Resing et al., 2017; Elliott et al., 2018; Resing et al., 2020) found that meaningful learning gains occur mainly when participants receive structured prompts that help them grasp the underlying problem-solving principles. In contrast, the absence of such support, as in the present design, shows the lower bound of what can be captured through repetition alone—practice-related improvements that do not necessarily reflect broader cognitive modifiability. Consistent with this, Hausknecht et al. (2007) demonstrated that performance gains are particularly pronounced when identical test forms are administered repeatedly, whereas improvements are considerably smaller with equivalent forms. It further demonstrates that, without feedback or instruction, improvements remain item-specific, short-lived and do not reflect LP. Future research should therefore examine how different levels of

instructional support influence test learning and where the transition from simple practice effects to genuine LP begins.

Limitations

Several limitations of the present study need to be considered. First, the short retest interval of only two weeks likely contributed to the magnitude of the observed practice effects. Prior meta-analytic research has shown that retest gains tend to decrease as the interval between test administrations increases, because memory for specific items fades and strategic recall becomes less effective over time (Hausknecht et al., 2007). As such, the strong improvements found for identical test versions in this study may partly reflect the recency of exposure which limits the generalizability of these findings to longer time spans that are more typical in educational or clinical contexts.

Potential ceiling effects, particularly in the SPM, must be taken into consideration: The SPM are known to provide limited differentiation at the higher end of the ability distribution, especially in university student samples. This restriction of measurement range may have attenuated the ability to detect subtle learning gains among high-performing individuals and could also have underestimated the variability in practice effects across the sample. At T1, the mean proportion of correctly solved items (i.e., relative solution probability) was already high (M = .87, SD ≈ .13, range = .44–1.00), with a quarter of the sample achieving perfect scores. This pattern indicates a ceiling effect, which likely limited variability as well as reduced internal consistency (α = .47). Consistent with this interpretation, mean performance at T2 (identical and equivalent) remained similarly high (M = .80–.88), which suggests that many participants were already close to the maximum possible score at baseline. Ceiling effects are common in shortened or less challenging subsets of the SPM when administered to high-performing samples. In contrast, the inclusion of the APM, which offers a broader difficulty range, partly addressed this issue but also emphasized that practice effects may differ across instruments depending on their psychometric properties.

Another limitation concerns the relatively low internal consistencies observed for the administered tests. The shortened versions of the SPM and APM, in particular, showed low reliability, which reduces measurement precision and may have attenuated the true magnitude of performance changes. In addition, the WST yielded only moderate internal consistency (α = .68) and MWT-A yielded questionable internal consistency (r = .68), suggesting that even for a standardized vocabulary measure, random error variance may have obscured smaller effects. In consequence, true score differences might have been underestimated.

The comparability of the verbal tests poses another limitation. Although the MWT-A was chosen as an equivalent measure to the WST, the two instruments are not psychometrically identical. This discrepancy in item difficulty complicates the interpretation of results. The observed performance decline on the MWT-A at T2 might therefore not reflect a genuine negative LP but rather differences in difficulty. Consequently, it is difficult to draw firm conclusions about LP in the verbal domain when equivalent versions vary systematically in difficulty. What appears as a change in ability might simply be due to differences in the tests themselves. Without proper equivalence between versions, it is impossible to determine whether changes in performance reflect actual learning, test fatigue, or differences in item difficulty. This underlines the importance of carefully matching test forms when investigating LP, because deviations can lead to ambiguous interpretations.

Furthermore, the MWT-B has been developed as an equivalent form to the MWT-A (Lehrl et al., 1977). A more suitable approach might be to use both versions in a test–retest design because it would provide a more reliable and psychometrically sound way to estimate LP rather than the WST.

A methodological strength of this proof-of-concept design is the homogeneity of the sample, which ensured comparability and reduced confounding influences. Nevertheless, the targeted recruitment of psychology students naturally limits the generalizability of the findings to more diverse populations.

Contribution

The present study provides two key insights. The results show that tests measuring the same cognitive domain are not a straightforward solution to control for practice effects. It also showed that the difficulty of alternative versions can vary significantly, which leads to new interpretative challenges. Another point worth stating is that the study situates practice effects within the broader framework of DT. Mere repetition, in the absence of explicit feedback or instructional support, is insufficient to produce the kind of adaptive and transferable learning processes described in Guthke's (1996) conception of DT. While participants showed improvements when repeating the identical test version, these gains did not generalize to the comparable form, indicating that the observed effects likely reflect practice or familiarity rather than genuine LP. This aligns with Guthke's (1996) idea that DT aims to engage test-takers actively in the learning process, using feedback as a tool to guide and enhance their performance.

The present results therefore provide tentative support for Guthke's (1996) framework, suggesting that while simple retesting can elicit limited performance gains, the core

mechanisms of DT—such as facilitation and guided feedback—are essential for revealing genuine LP.

The findings also suggest directions for future work: It would be useful to compare DT more directly with newer approaches such as adaptive testing, or with status-oriented procedures that include feedback without explicit instruction (Guthke et al., 1996). This could help to clarify which parts of the testing process promote learning and which only reflect familiarity.

Conclusion

The goal of the present study was to to investigate whether simple test repetition, without any instructional support or feedback, can be considered a learning intervention within the framework of DT. Across domains, participants exhibited significant performance gains when retested with identical test versions, whereas equivalent versions produced little to no improvement. These findings reveal that repetition primarily simplifies item-specific familiarity and recall but it does not provide evidence of transferable LP.

This study supports in achieving a more accurate comprehension of practice effects in cognitive testing. It investigated performance improvements that stemmed from repetition. Participants showed improvement primarily when they repeated the exact same test version, but these gains did not extend to the comparable version. This demonstrates that the adaptive and transferable learning processes highlighted in DT (Guthke & Wiedl, 1996; Tzuriel, 2001) require more than just repeated exposure. The results thus contribute to distinguishing practice effects from genuine cognitive modifiability. It further emphasizes the importance of incorporating supportive elements in DT to reveal LP in the first place.

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General,* 117(3), 288–318. https://doi.org/10.1037/0096-3445.117.3.288

Beckmann, J. F. (2004). Nachruf auf Jürgen Guthke. *Zeitschrift für Differentielle und Diagnostische Psychologie,* 25(2), 117–119. https://doi.org/10.1024/0170-1789.25.2.117

Boosman, H., Bovend'Eerdt, T., Visser-Meily, J., Nijboer, T. & van Heugten, C. (2016). Dynamic testing of learning potential in adults with cognitive impairments: A systematic review of methodology and predictive value. *Journal of Neuropsychology* 10(2), 186-210. https://doi.org/10.1111/jnp.12063

Börnert-Ringleb, M. (2018). Die Erfassung von Lernpotentialen und benötigter Unterstützung: Dynamisches Testen. *Potsdamer Zentrum für empirische Inklusionsforschung (ZEIF),* 2018, Nr. 4. https://doi.org/10.25932/publishup-67206

Carpenter, P., Just, M. & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological review,* 97(3), 404. https://doi.org/10.1037/0033-295X.97.3.404

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1–22. https://doi.org/10.1037/h0046743

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin,* 132(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Champely S (2020). *pwr: Basic Functions for Power Analysis.* R package version 1.3-0, https://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical Power Analysis for the Behavior Sciences (2nd Edition).* Routledge.

Elliott J., Resing W. & Beckmann J. (2018). Dynamic assessment: A case of unfulfilled potential? *Educational Review*, 70(1), 7–17. https://doi.org/10.1080/00131911.2018.1396806

Ellis, J.L. (2013). A standard for test reliability in group research. *Behavioral Research* (45), 16–24. https://doi.org/10.3758/s13428-012-0223-z

Feuerstein, R. (1980). *Instrumental Enrichment: An intervention program for cognitive modifiability.* University Park Press.

Flanagan, D. P., & Harrison, P. L. (Eds.). (2012). *Contemporary intellectual assessment: Theories, tests, and issues (3rd ed.).* The Guilford Press.

George, D., & Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference*. 11.0 Update (4th ed.). Allyn & Bacon.

Grigorenko, E. & Sternberg, R. (1998). Dynamic testing. *Psychological Bulletin,* 124(1), 75-111. https://doi.org/10.1037/0033-2909.124.1.75

Grigorenko, E. & Sternberg, R. (2002). Dynamic Testing. The Nature and Measurement of Learning Potential. Cambridge University Press.

Guthke, J. (1992). *Learning tests-the concept, main research findings, problems and trends. Learning and Individual Differences, 4(2), 137–151.* https://doi.org/10.1016/1041-6080(92)90010-C

Guthke, J. (Hrsg.). (1995). *Adaptive computergestützte Intelligenzlerntestbatterie (ACIL*). Schuhfried.

Guthke, J., Beckmann, J. & Wiedl, K. (2003). Dynamik im Dynamischen Testen [Dynamics in dynamic testing]. *Psychologische Rundschau, 54* (4), 225–232. https://doi.org/10.1026//0033-3042.54.4.225

Guthke, J. & Wiedl, K. (1996). *Dynamisches Testen. Zur Psychodiagnostik der intraindividuellen Variabilität.* Hogrefe.

Hammers, D., Bothra, S., Polsinelli, A., Apostolova, L., & Duff, K. (2024). Evaluating practice, effects across learning trials – ceiling effects or something more? *Journal of Clinical and Experimental Neuropsychology* 46, 630 - 643. https://doi.org/10.1080/13803395.2024.2400107

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Haywood, H. & Tzuriel, D. (2002). Applications and Challenges in Dynamic Assessment. *Peabody Journal of Education*, 77(2), 40-63. https://doi.org/10.1207/S15327930PJE7702_5

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children (KABC, K-ABC)* [Database record]. APA PsycTests. https://doi.org/10.1037/t27677-000

Kubinger, K. (2019). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens.* Hogrefe.

Lehrl, S. (1977). *Der Mehrfachwahlwortschatztest MWT-B.* Straube.

Lehrl, S., Daun, H. & Schmidt, R. (1971). Eine Abwandlung des HAWIE-Wortschatztests als Kurztest zur Messung der Intelligenz Erwachsener. *Archiv für Psychiatrie und Nervenkrankheiten,* 215, 353-364. https://doi.org/10.1007/BF00342670

Lehrl, S., Merz, J., Erzigkeit, H., & Galster, V. (1974). Der MWT-A — ein wiederholbarer Intelligenz-Kurztest, der weitgehend unabhängig von seelisch-geistigen Störungen ist. [MWT-A--a repeatable intelligence short-test, fairly independent from psychomental disorders]. *Der Nervenarzt,* 45(7), 364–369.

Lehrl, S., Merz, J., Burkhard, G., & Fischer, S. (1991). *Mehrfachwahl-Wortschatz-Intelligenztest.* Spitta.

Lidz, C. (1987). *Dynamic Assessment: An interactional approach to evaluating learning potential.* Guilford.

Luria, A. (1976). *Cognitive development: Its cultural and social foundations*. Harvard University press.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology,* 49(2), 270–291. https://doi.org/10.1037/a0028228

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist,* 51(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

Nurhudaya, N., Taufik, A., Yudha, E. & Suryana, D. (2019). The Raven's Advanced Progressive Matrices in Education Assessment with a Rasch Analysis. Universal Journal of Educational Research,7(9),1996-2002. https://doi.org/10.13189/ujer.2019.070921

Pallentin, V., Danner, D., Lesche, S. & Rummel, J. (2024). Validation of the Short Parallel and Extra-Short Form of the Heidelberg Figural Matrices Test (HeiQ). *Journal of Intelligence.* 12(10), 100. https://doi.org/10.3390/jintelligence12100100

Pind, J., Gunnarsdóttir, E., & Jóhannesson, H. (2003). Raven's Standard Progressive Matrices: new school age norms and a study of the test's validity. *Personality and Individual Differences*, 34(3), 375-386. https://doi.org/10.1016/S0191-8869(02)00058-2

Qualtrics (2024). *Qualtrics* [Computer software]. Qualtrics. https://www.qualtrics.com

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Raven, J. (2009). *Standard Progressive Matrices (SPM).* Pearson.

Raven, J. (1998). *Advanced Progressive Matrices (APM).* Pearson.

Reber, A. S. (1993). *Implicit learning and tacit knowledge.* Oxford University Press.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3

Resing, W., Elliott, J., & Vogelaar, B. (2020). Assessing Potential for Learning in School Children. *Oxford Research Encyclopedia of Education.* https://doi.org/10.1093/acrefore/9780190264093.013.943

Resing, W., Touw, K., Veerbeek, J. & Elliott, J. (2017). Progress in the inductive strategy-use of children from different ethnic backgrounds: a study employing dynamic testing, *Educational Psychology,* 37(2), 173-191. https://doi.org/10.1080/01443410.2016.1164300

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, 22(6), 800–811. https://doi.org/10.1037/a0013091

Sbaibi, R., Aboussaleh, Y., & Ahami, A. (2014). The Standard Progressive Matrices Norms in an international context among the middle school children of the rural commune Sidi el Kamel (North-Western Morocco). *Web Psych Empiricist*, 7(28), 1-13.

Schmidt, K., Mattis, P., Adams, J. & Nestor, P. (2005). Alternate-form reliability of the Dementia Rating Scale-2. *Archives of Clinical Neuropsychology*, 20(4), 435-441. https://doi.org/10.1016/j.acn.2004.09.011

Schmidt, K. & Metzler, P. (1992). *Wortschatztest*. Beltz Test.

Schunk, D. H. (2012). *Learning theories: An educational perspective* (6th ed.). Pearson Higher Ed.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *American Journal of Psychology,* 15(2), 201–293. https://doi.org/10.2307/1412107

Stevenson, C., Heiser, W. & Resing, W. (2016). Dynamic testing: Assessing cognitive potential of children with culturally diverse backgrounds. *Learning and Individual Differences*, 47, 27-36. https://doi.org/10.1016/j.lindif.2015.12.025

Sun, R., Slusarz, P., & Terry, C. (2005). The Interaction of the Explicit and the Implicit in Skill Learning: A Dual-Process Approach. *Psychological Review*, 112(1), 159–192. https://doi.org/10.1037/0033-295X.112.1.159

Thorndike, E. (1924). *An introduction to the theory of mental and social measurement.* Wiley.

Thurstone, L.L. (1938). *Primary mental abilities.* University of Chicago Press.

Tiekstra, M., Minnaert, A. & Hessels, M. (2016). A review scrutinising the consequential validity of dynamic assessment. *Educational Psychology*, 36(1), 112-137. https://doi.org/10.1080/01443410.2014.915930

Touw, K., Vogelaar, B., Thissen, F., Rovers, S., & Resing, W. (2019). Progression and individual differences in children's series completion after dynamic testing. *British Journal of Educational Psychology*. 90, 184-205.https://doi.org/10.1111/bjep.12272

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, 247(4940), 301–306. https://doi.org/10.1126/science.2296719

Tzuriel, D. (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review*, *12*, 385-435.

University of Vienna. (n.d.). Laboratory Administration for Behavioral Sciences (LABS). https://labs-univie.sona-systems.com/Default.aspx?ReturnUrl=%2f

Veerbeek, J., & Vogelaar, B. (2024). Dynamic Testing of Instructional Needs: A Training-Only Graduated Prompts Approach. *Journal of Psychoeducational Assessment*, 43(1), 74-87. https://doi.org/10.1177/07342829241287947

Vo, D. & Csapó, B. (2022). Measuring inductive reasoning in school contexts: a review of instruments and predictors. International Journal of Innovation and Learning, 31(4), 506-525. https://doi.org/10.1504/IJIL.2022.123179

Vogelaar, B., Veerbeek, J., Splinter, S. E., & Resing, W. C. (2021). Computerized dynamic testing of children's potential for reasoning by analogy: The role of executive functioning. *Journal of Computer Assisted Learning,* 37(3), 632-644. https://doi.org/10.1111/jcal.12512

Wickham H, Bryan J (2023). *readxl: Read Excel Files.* R package version 1.4.3, https://CRAN.R-project.org/package=readxl

Wechsler, D. (1944). *The measurement of adult intelligence (3rd ed.).* Williams & Wilkins Co. https://doi.org/10.1037/11329-000

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV).* Pearson.

Wiedl, K. (2003). Dynamic testing: A comprehensive model and current fields of application. Journal of Cognitive Education and Psychology, 3(1), 93–119. https://doi.org/10.1891/194589503787383055

Wood, E., Biggs, K., & Molnar, M. (2024). Characteristics of Dynamic Assessments of Word Reading Skills and Their Implications for Validity: A Systematic Review and Meta-analysis. Sage Open, 14(4). https://doi.org/10.1177/21582440241300536

Yip, C. C. H., Ouyang, X., & Zhang, X. (2025). Measuring general cognitive ability in early elementary students: A shortened version of the Raven's Standard Progressive Matrices. *Intelligence*, 112, 101949. https://doi.org/10.1016/j.intell.2025.101949

Zaaiman, H., Van Der Flier, H., & Thijs, G. D. (2001). Dynamic testing in selection for an educational programme: Assessing South African performance on the Raven Progressive Matrices. International Journal of Selection and Assessment, 9(3), 258-269. https://doi.org/10.1111/1468-2389.00178

Zubin, J. (1950). Symposium on statistics for the clinician. *Journal of Clinical Psychology*, 6(1), 1–6. https://doi.org/10.1002/1097-4679195001

List of tables

Table 1 17
Table 2 24
Table 3 25

List of Figures

Figure 1.. 19
Figure 2.. 20
Figure 3. 21
Figure 4.. 23

List of Abbrevations

ANCOVA	Analysis of covariance
APM	Advanced Progressive Matrices
CI	Confidence interval
DA	Dynamic Assessment
DT	Dynamic Testing
LP	Learning Potential
MWT-A	Mehrfachauswahl-Wortschatztest, Version A
MWT-B	Mehrfachauswahl-Wortschatztest, Version B
SCM	Theory of Structural Cognitive Modifiability
SE	Standard error
SPM	Standard Progressive Matrices
T1	First testing session
T2	Second testing session
WST	Wortschatztest
ZPD	Zone of Proximal Development

Abstract

This study examined whether simple test repetition, without instructional support, can be considered a learning intervention within the framework *of dynamic testing* (DT). DT, specifically learning tests as postulated by Guthke, typically combines pretest, training, and posttest phases to assess learning potential. In contrast, the present proof-of-concept study aimed to isolate repetition as the most minimal intervention and therefore evaluate its effects on cognitive performance. A within-subjects design with repeated measures was employed. At the first measurement point, all participants (N = 163 psychology students from the University of Vienna) completed items from the *Standard Progressive Matrices* (SPM), *Advanced Progressive Matrices* (APM), and the *Wortschatztest* (WST). At the second measurement point, participants were randomly assigned to either identical or equivalent versions of these tasks. Performance differences between the first and second measurement point were analyzed using paired-samples t-tests, Wilcoxon signed-rank tests, and ANCOVA. Results revealed significant practice effects for identical versions of the SPM, APM, and WST, while comparable versions showed little or no improvement. The lower T2 scores observed in the MWT-A, compared with the WST, most likely reflect differences in item difficulty rather than genuine performance decline. Taken together, the results point toward Guthke and Wiedl's (1996) framework, suggesting that participants with higher verbal ability tended to benefit somewhat more from retesting. This pattern can be seen as a cautious indication of individual modifiability, even though the effects remain modest. At the same time, the data confirm that repeated testing mainly leads to task-specific practice gains rather than broader, transferable learning effects. Implications for the assessment of learning potential and directions for future research are discussed.

Zusammenfassung

Die vorliegende Studie untersuchte, ob reine Testwiederholung ohne Instruktion oder Feedback als Lernintervention im Rahmen des dynamischen Testens betrachtet werden kann. *Dynamisches Testen* (DT) umfasst typischerweise Pretest-, Trainings- und Posttestphasen zur Erfassung des *Lernpotenzials* (LP). In Abgrenzung dazu zielte diese Proof-of-Concept-Studie darauf ab, Wiederholung als minimale Form der Intervention zu isolieren und somit deren Effekte auf kognitive Leistungen zu prüfen. Um dies zu untersuchen, wurde ein Within-Subjects-Design mit Messwiederholung eingesetzt. Zum ersten Messzeitpunkt bearbeiteten alle Teilnehmenden (N = 163 Psychologiestudierende der Universität Wien) Aufgaben aus den *Standard Progressive Matrices* (SPM), den *Advanced Progressive Matrices* (APM) sowie dem *Wortschatztest* (WST). Am zweiten Messzeitpunkt wurden sie zufällig entweder identischen oder äquivalenten Testversionen zugewiesen. Die Differenzen zwischen dem ersten und zweiten Erhebungszeitpunkt wurden mittels gepaarter t-Tests, Wilcoxon-Tests und ANCOVA ausgewertet. Die Ergebnisse zeigten signifikante Übungseffekte für die identischen Versionen der SPM, APM und des WST, während die äquivalenten Versionen nur geringe oder keine Verbesserungen aufwiesen. Die niedrigeren T2-Werte im MWT-A im Vergleich zum WST lassen sich höchstwahrscheinlich auf Unterschiede in der Itemschwierigkeit und nicht auf einen tatsächlichen Leistungsrückgang zurückführen. Insgesamt deuten die Befunde auf das Rahmenmodell von Guthke und Wiedl (1996) hin, da Teilnehmende mit höherer verbaler Fähigkeit tendenziell stärker von der Wiederholung profitierten. Dieses Muster kann als vorsichtiger Hinweis auf individuelle Modifizierbarkeit interpretiert werden, wenngleich die Effekte gering blieben. Gleichzeitig bestätigen die Daten, dass reine Testwiederholung vor allem aufgabenspezifische Übungseffekte hervorruft und keine breiteren, transferierbaren Lernzuwächse bewirkt. Die Ergebnisse werden im Hinblick auf Implikationen für die Erfassung von Lernpotenzial und zukünftige Forschungsrichtungen diskutiert.

Erklärung zur Verwendung von Künstlicher Intelligenz (KI)

Hiermit erkläre ich, dass ich die folgenden KI-Hilfsmittel zur Erstellung meiner Masterarbeit mit dem Titel „Revisiting Guthke's process-oriented concept of psychological assessment: A proof-of-concept study in longitudinal data" verwendet habe. Ich versichere, dass die Kennzeichnung der verwendeten KI-Hilfsmittel vollständig ist, inkl. Verwendungszweck, Produktname, als auch ggfs. Prompts und Outputs des KI-Hilfsmittels. Ich verantworte die Auswahl und die Übernahme sämtlicher Ergebnisse der von mir verwendeten KI-Hilfsmittel vollumfänglich selbst.

Verwendungszweck*	**Produktname des KI-Hilfsmittels**	**Ggfs. Prompt (Input) und Output des KI-Hilfsmittels**
Einzelne Übersetzungen, teilweise übernommen	Deepl	
Formulierungshilfen, teilweise übernommen	Scribbr Text umschreiben	
Statistische Empfehlungshilfen	Statsbot (Universität Wien)	A1, A2, A3, A4, A5

* Hilfsmittel zur Überprüfung von Grammatik und Rechtschreibung oder der automatisierten Erstellung des Literaturverzeichnisses sind von dieser Erklärung ausgenommen.

Name: Lisa Kasal

Matrikelnummer: 11925212

Ort, Datum: Wien, 28.10.2025
Unterschrift: Lisa Kasal

Anhang

A1 StatsBot
Hello Lisa Marie, I am your Statistics tutor. What question do you have?

A1
hi! i wrote this in my masters thesis: To test Hypotheses 1a and 1b, six one sample t-tests were conducted separately for the identical and equivalent test versions using difference scores between T1 and T2 to examine whether performance significantly improved from T1 to T2. Individual difference scores be-tween T1 and T2 were computed for each participant, and it was examined whether these mean differences significantly deviated from zero. For Hypothesis 2, the difference scores of the identical and equivalent versions were directly compared using pairedsamples t-tests. my professor wrote: That can't be right. One-sample t-tests are used to calculate potential differences between a mean value and a theoretical value (e.g., 0). That's not what it says in the pre-registration either. It specifies paired-sample t-tests of pre- vs. post-test scores (which would make sense). So, unlike stated above, the delta variable here is not the AV either. what should I do now?

StatsBot - Output
Your professor's feedback suggests there's a misunderstanding or error in the description of your statistical analysis, particularly regarding the type of t-tests you used or intended to use. Here are steps you can take to address this: 1. **Clarify Your Analysis Plan**: - Revisit your pre-registration document and confirm what type of t-tests you originally planned to use for your analyses. This will help you understand the discrepancy between your thesis and the pre-registration. 2. **Correct the Statistical Approach**: - If your intent was to compare pre-test and post-test scores, paired-samples t-tests are appropriate because they assess whether the means of two related groups (e.g., the same participants at two different times) differ. - If you mistakenly described the use of one-sample t-tests, revise your thesis to indicate paired-samples t-tests were used to test Hypotheses 1a and 1b. 3. **Revise Your Thesis**: - Update the relevant sections in your thesis to accurately reflect the statistical tests you used. For Hypotheses 1a and 1b, describe the paired-samples t-tests comparing T1 and T2 scores. - Ensure your explanation aligns with the pre-registration and accurately describes the analysis conducted. 4. **Consult with Your Professor**: - Discuss the revisions with your professor to ensure your updated explanation aligns with their expectations and the intended analysis. 5. **Update Your Results**: - Depending on the changes you make, verify that your results section accurately reflects the statistical tests used and the findings. Ensure all interpretations are consistent with the correct analysis method. By clarifying and correcting your statistical analysis and aligning it with your pre-registration, you'll be able to address your professor's concerns effectively.

A2

but i computed the difference scores T2-T1 and then I used a one sample t-test. so technically it is a test of differences?

StatsBot - Output

Your approach of computing difference scores (T2 - T1) and then using a one-sample t-test is indeed a valid way to test whether the average change in scores is significantly different from zero. Here's a clarification on how this method works and how you might address your professor's feedback: ### One-Sample T-Test on Difference Scores: - **Methodology**: By calculating the difference scores for each participant (T2 - T1), you create a new variable that represents the change or improvement from T1 to T2. - **Analysis**: You then conduct a one-sample t-test on these difference scores to determine if their mean deviates significantly from zero. This approach tests whether there is a significant average change in scores from T1 to T2 across participants. ### Addressing Your Professor's Feedback: 1. **Clarify Your Approach**: - Explain to your professor that you used the difference scores to simplify the analysis to a one-sample t-test, which is a common practice for examining pre- and post-test changes. - Highlight that while the test is technically a one-sample t-test, it is assessing the change between two time points by testing if the average difference score is significantly different from zero. 2. **Align with Pre-Registration**: - Confirm that this approach is consistent with your pre-registration, or adjust your thesis to align with what was specified. If the pre-registration specified paired-samples t-tests, you might need to reframe your analysis description to match. 3. **Revise Your Thesis Description**: - Update the thesis to clearly describe the rationale behind using one-sample t-tests on difference scores, emphasizing that this method effectively tests for significant changes between T1 and T2. 4. **Consult Your Professor**: - Discuss this with your professor to ensure they understand your rationale and confirm whether the one-sample t-test approach on difference scores is acceptable. By clearly explaining your approach and ensuring it aligns with your pre-registration document, you can address the misunderstanding and make necessary revisions to your thesis.

A3

ok. just to make sure: my intent with this approach was to compute the difference of each individual. does it make a statistical difference if i a) compute the difference of T2 to T1 and then calculate a one sample t test or b) use a paired t-test for T2 to T1?

StatsBot Output

The statistical intent of both approaches is similar: to assess whether there is a significant change from T1 to T2. However, the methods are mathematically equivalent, and the choice between them often comes down to convention or clarity in reporting. Here’s how they compare: ### Option A: One-Sample T-Test on Difference Scores - **Method**: Calculate the difference for each participant (T2 - T1), then conduct a one-sample t-test to see if the mean of these differences significantly deviates from zero. - **Interpretation**: This

tests whether, on average, the change from T1 to T2 is significantly different from zero. - **Use**: This method is straightforward and effectively tests for average change. It is mathematically equivalent to the paired samples t-test. ### Option B: Paired-Samples T-Test - **Method**: Directly compare the two sets of scores (T1 and T2) using a paired-samples t-test, which inherently calculates the difference between paired observations (similar to difference scores) and tests if the average difference is zero. - **Interpretation**: This tests the same hypothesis as the one-sample t-test on difference scores, but is often preferred for clarity in reporting paired data since it directly indicates that you are comparing two related groups. - **Use**: This is the conventional method for analyzing pre- and post-test data because it explicitly acknowledges the paired nature of the data. ### Conclusion: - **Statistical Equivalence**: Both methods are statistically equivalent in terms of results when applied correctly. They will yield the same p-value and conclusions about the significance of the change. - **Recommendation**: If your pre-registration or conventions in your field suggest using paired-samples t-tests, it may be beneficial to frame your analysis in that context for consistency and clarity. If you've already computed difference scores, explaining that your approach is mathematically equivalent to the paired-samples t-test can clarify your intent. Ultimately, either approach will accurately test for significant changes between T1 and T2, but using the paired-samples t-test is often clearer in communicating the analysis of pared observations.

A4 what is your recommendation if i have already calculated the one samples and this was my professors feedback?
Output
Given your situation, where you've already calculated the one-sample t-tests and received feedback from your professor, here are steps you can take to address this effectively: 1. **Understand the Feedback**: - Your professor seems to expect paired-samples t-tests based on your pre-registration. It's important to align your analysis with the expectations laid out before your study began. 2. **Explain Your Approach**: - Provide a brief explanation to your professor that the one-sample t-test on difference scores is mathematically equivalent to the paired-samples t-test. Highlight that both tests assess whether the average change (T2 - T1) is significantly different from zero. 3. **Revise Your Thesis**: - Consider reframing your analysis in terms of paired-samples t-tests, which might align better with academic conventions and your pre-registration. You can mention that you computed difference scores as an intermediate step, but the statistical test is equivalent. 4. **Update Your Documentation**: - Ensure that your thesis clearly describes the statistical test used and aligns with your pre-registration. If using the paired-samples t-test is expected, update your results section accordingly to reflect this test. 5. **Consult Your Professor**: - Discuss your explanation and proposed revisions with your professor to ensure they agree with your approach and are satisfied with the changes. 6. **Consider Re-Running the Analysis (Optional)**: - If you want to align strictly with conventions or your pre-registration, you might consider re-running the analysis using paired samples t-tests. Given that the results

will be the same, this is more about meeting expectations than altering findings. By addressing the feedback in a way that aligns with academic conventions and your pre-registration, you'll likely resolve the issue to your professor's satisfaction.

A5
do i get the exact same results ?

StatsBot Output

Yes, you should get the exact same results using either method, assuming no computational errors. Both the one-sample t-test on difference scores and the paired-samples t-test are mathematically equivalent when applied to paired data like pre-test and post-test scores. Here's why: - **One-Sample T-Test on Difference Scores**: This test involves calculating the difference for each participant (T2 - T1) and then examining if the average of these differences significantly deviates from zero. - **Paired-Samples T-Test**: This test directly compares the two sets of scores (T1 and T2) for each participant and assesses whether the average difference between them is significantly different from zero. In both cases, you're essentially evaluating the mean difference between two related sets of scores, so the t-statistic, degrees of freedom, and p-value will be identical. This equivalence arises because the paired-samples t-test inherently calculates the difference between paired observations, just as you did manually with the difference scores. Thus, you can confidently convey to your professor that the statistical findings (such as the significance level and test statistics) will be the same regardless of which method you use, although the paired-samples t-test is often preferred for clarity in reporting paired data analyses