International Journal of Educational Research 117 (2023) 102097

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

Educational inequality due to lack of validity: A methodological critique of the Dutch school system

S. Scheider [a,*], S. Rosenfeld [b], S. Bink [b,c], N. Lecina [c]

[a] Department of Human Geography and Spatial Planning, Utrecht University, NL, Netherlands
[b] Independent teacher and researcher, Utrecht, NL, Netherlands
[c] Studio Moio, Leiden, NL, Netherlands

ARTICLE INFO

Keywords:
Equality of chances
Validity
Assessment
Teaching paradigms
Dutch school policy

ABSTRACT

The Netherlands have a problem regarding quality as well as equality in their school system. Many students fail to reach minimal skill levels and underachieve with respect to their own learning capacities. They end up on educational levels that do not correspond to their learning potential, especially when their parents do not speak the language or cannot afford tutoring to exploit this potential. In this article, we argue that important reasons for this development lie in a lack of validity of judgments on various levels of the educational system. We summarize these reasons into 8 theses about the validity of methods, starting from methods of testing and assessment, over the design of the teaching process and underlying paradigms, and ending with the more general cybernetic effects of school policy and the meritocratic social paradigms in which they are embedded.

1. Introduction

Similar to other European countries, the Netherlands ratified international laws which *guarantee equal access* to educational levels for every child, based on individual merits[1]. Yet, there is an increasing awareness in the Dutch society that the Dutch school system has a problem concerning equality[2] as well as quality. According to the Programme for International Student Assessment (PISA), the quality of the Dutch school system (which used to be outstanding) has been on a decline since 2006[3]. School inspectors are warning that reading standards, as well as arithmetic, math and science skills, have dropped considerably during the last 20 years. A report of the Dutch *inspectorate of education* ('Onderwijsinspectie') from 2019 has discovered that the equality of chances has likewise corroded considerably during the same period (Baltussen, 2019), and the newest edition comes to similar conclusions (Oppers, 2022).

One reason is that children whose parents cannot prepare their kids for tests or pay for private tutoring do not have the same chance of making educational progress (Marks et al., 2006), even if they have sufficient learning potential. Dutch parents who know how the

Abbreviations: CITO, centraal instituut voor toetsontwikkeling; LVS, leerlingvolgsysteem; vwo, voorbereidend wetenschappelijk onderwijs; havo, hoger algemeen voortgezet onderwijs; vmbo, voorbereidend middelbaar beroepsonderwijs.

* Corresponding author.
E-mail address: s.scheider@uu.nl (S. Scheider).

[1] According to, e.g., the European convention on Human Rights (ECHR) (Pillai, 2012) Article 2 of Protocol Number 1, as well as the UN universal declaration of human rights Article 26 (United Nations, 1948).

[2] Equality understood here in terms of accessibility and horizontality, cf. (McCowan, 2016).

[3] https://www2.compareyourcountry.org/pisa/country/nld?lg=en.

https://doi.org/10.1016/j.ijer.2022.102097
Received 20 May 2022; Received in revised form 5 October 2022; Accepted 11 November 2022
Available online 28 November 2022

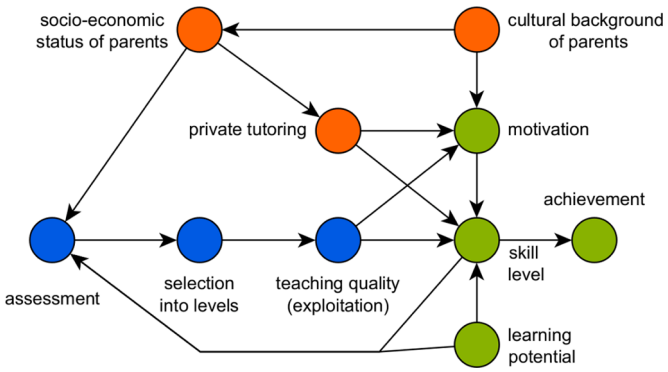


Fig. 1. Causal diagram of educational inequality. Red factors are related to social background, blue factors capture the quality of a school system, and green factors are characteristics of a child. To reduce inequality, the role of the background in exploiting learning potentials needs to be reduced.

system works, on the other hand, can use it to their advantage (Elffers, 2018), by exploiting the learning potential of their children from the privileged position of their own social background (orange factors in Fig. 1). However, parents are not to blame. The only way to mitigate this effect on inequality is to compensate for unequal background conditions by sound *assessment, selection and teaching methods* in publicly financed schools (blue factors in Fig. 1). This is supported by multifarious empirical evidence. For example, studies based on PISA[4] data have revealed that early and institutionalized student selection into levels amplifies the effect of cultural and socio-economic background on achievements (Dupriez and Dumay, 2006; Gorard and Smith, 2004; Marks et al., 2006). Furthermore, it is well known that the qualities of teaching are crucial for mitigating such effects on inequality (Hanselman, 2018), and that early student assessment and *tracking* methods[5] have a significantly negative effect on educational achievements, as shown by a recent meta-study (Felouzis and Charmillot, 2013; Gamoran and Mare, 1989; Terrin and Triventi, 2022). While previous research has focused on comparative educational statistics for investigating such causes of inequality (cf. Felouzis and Charmillot (2013); Yang Hansen and Strietholt (2018)), studies on the role of the *methodological validity* of assessments for educational inequality, especially in a Dutch national context, are lacking. In this article, we argue that an important reason for the existing inequality problem in the Netherlands is a *lack of validity of educational methods* with respect to assessment, selection and teaching.

In our everyday lives, we constantly rely on the judgment of other people who take on the responsibility of deciding in our names about issues we cannot decide ourselves, be it judges, policemen, politicians, as well as teachers. For obvious reasons, our lives and biographies depend on whether such judgments are *valid*. Validity implies that *criteria are defined relative to specific goals* of what needs to be judged. A judgment is valid if it can be approved by everyone who accepts the criteria for this goal, including ourselves (Janich, 2001).

For example, an important form of judgment in school are *assessments of learning outcomes* (cf. Moss et al. (2006)). In this case, validity criteria are defined with respect to learning goals. Valid assessments measure whether pupils have acquired precisely the skills that correspond to such goals (McKeachie and Svinicki, 2013, ch. 10). However, misjudgment of skills can arise when confusing goals. For example, a test of math skills given in terms of a story ('verhaaltjessommen') may require language skills that intermix with the learning goals of algebra, leading to a misjudgment of algebra skills. Furthermore, to grant equal chances to children, valid assessments are required not only for learning outcomes, but also for *learning potentials* (Vogelzang, 2016). In this case, static tests are not a valid assessment method because we are targeting a *disposition to learn*, not a learning history. To assess potentials it is required to observe performances in *dynamic assessments* (Tzuriel, 2012). This can be done by observing the modifiability of learning behavior focusing on maximum learning progress and not on averages. To be valid, dynamic assessments therefore need to be embedded into the learning process. Validity in this case can be based on the *success of trials*. For example, if a student's academic potential was assessed as low but later on, the student graduates at a university, this renders the assessment invalid. Finally, validity can also be considered for methods used in an educational system that are not assessments. For example, teaching methods as well as methods of school policy become invalid with respect to the goal of equality of chances precisely when they are not effective in exploiting learning potentials. We therefore argue that equality of educational chances requires *validity* along at least three dimensions:

1. *Valid assessments of skill levels*[6], to be able to judge the level of proficiency a child has acquired with respect to a learning goal.
2. *Valid assessments of learning potentials.*[7], to be able to judge whether a child will be able to master certain educational pathways.

[4] OECD's Programme for International Student Assessment https://www.oecd.org/pisa.
[5] Dividing the student population into groups according to their achievements.
[6] This is sometimes called 'content validity' in assessment research (Moss et al., 2006).
[7] This is roughly equivalent to '(predictive) criterion validity' (Moss et al., 2006). We deviate from this terminology here because criteria are relevant for all kinds of validity, and assessing learning potentials implies special criteria.

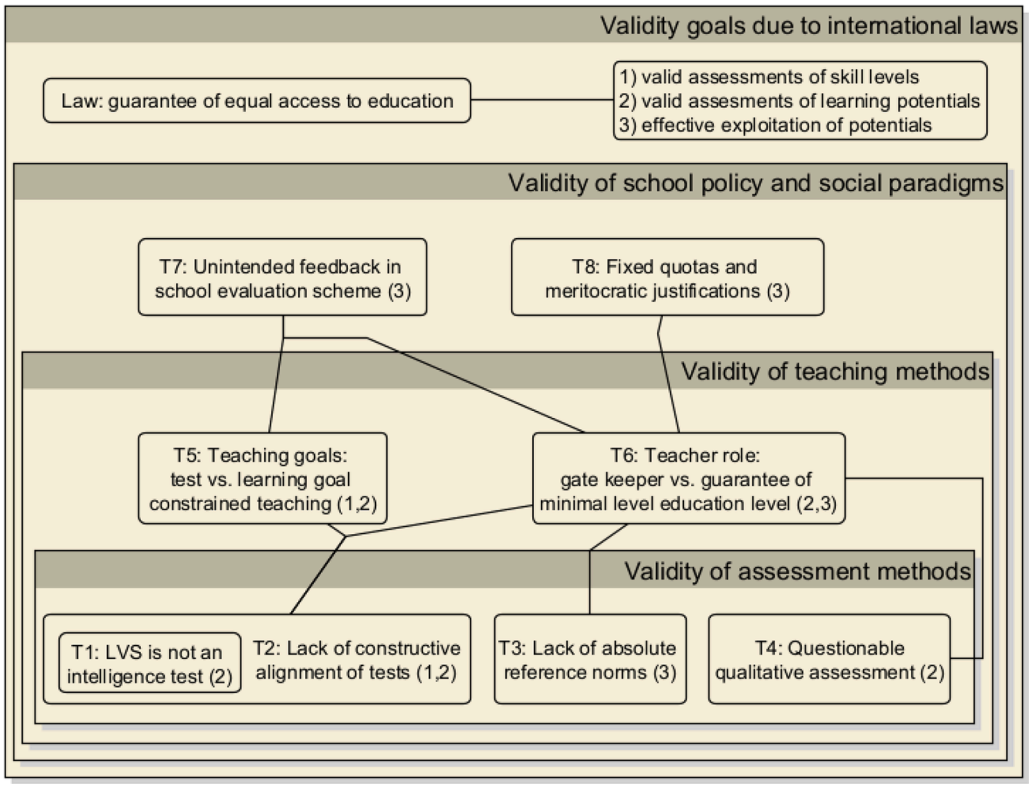


Fig. 2. Overview of the 8 theses substantiated in this article. They address the validity of educational methods on three nested layers of the educational system. Lines connect theses whose underlying methods are mutually dependent.

3. *Effective exploitation of these learning potentials* in teaching methods. Note the requirement to be effective, not efficient[8].

Yes, it is doubtful whether the Dutch system in primary and secondary school performs well along these three dimensions. Karen Heij, a researcher who has worked as an educational test designer, recently doubted the appropriateness of the Dutch selection and testing system (Heij, 2021). In our own critique, we will reuse some of her arguments, as well as refer to certain questions raised by Boonstra et al. (2020). In contrast to these authors, we explore reasons for a lack of validity of educational methods on different levels of the Dutch school system, and explain how these reasons tend to enforce each other. We substantiate our reasons in terms of *8 theses* which should be regarded as *hypotheses* requiring further research. As shown in Fig. 2, our theses address the question of validity on three nested levels: The level of assessment methods, the teaching methods, as well as the policy level.

2. Eight theses about a lack of validity as a reason for educational inequality in the Dutch school system

The following methodological critique is based on arguments from learning psychology, educational science and learning philosophy.

A *critique* (greek: the art of judgment) is a method in philosophy of using reason to check claims for their justification. It has its roots in Kant's critique of reason and in language philosophy (Janich, 2001), but has also been applied in psychology and education science, e.g., to critique the validity of PISA assessments (Fernandez-Cano, 2016; Wolf, 1998; Yang Hansen and Strietholt, 2018). In our case, we check whether validity assumptions underlying the current educational system are justified along the three dimensions mentioned above, based on the following sources:

- Statistical and experimental evidence from learning psychology and educational science
- Governmental reports, laws, as well as reports on educational history
- Test samples and other published educational tools
- Evidence from newspapers and online publications

Thus, our argumentation is based solely on available sources of knowledge and evidence, as referenced underneath each thesis.

[8] Being effective means to exploit as much as possible of a given potential. Being efficient means to exploit a given potential with minimal effort.

2.1. Assessment and selection into 'niveaus'

To decide about future paths on educational levels ('niveaus'), *state-wide standardized tests* are held regularly as part of a *student tracking system* ('leerlingvolgsysteem' (LVS)) (cf. Appendix A.1). In addition, certain qualitative forms of assessment are frequently used by teachers. In the following, we argue why underlying assessments are of questionable validity both as a means to assess skill levels as well as learning potentials.

We focus on the test developed by the Centraal instituut voor toetsontwikkeling (CITO)[9] as an example for LVS, which is most prominent. Yet, similar arguments can be made for other LVS systems, too.

2.1.1. CITO does not measure intelligence, it measures contingent skills

Parents as well as teachers in the Netherlands are frequently told that they should disregard and thus not 'interfere' with the CITO system. An often heard reason is that training could 'bias' the results of the test[10]. Training is unwanted because it could be used to 'manipulate' intelligence scores. Unfortunately, this view is not only based on a fundamental misconception of intelligence, it also misconceives CITO as a test of learning capacity.

IQ results consistently correlate with various other measures of intellectual capacity, including cognitive capabilities (memory and speed), learning capacity, as well as educational attainments (Plomin and Von Stumm, 2018). Yet, regarding the role of IQ tests in schools, two precautions are in order: First, IQ tests cannot measure learning progress and skill levels because they are designed to be independent of prior knowledge and trained skills. And second, though such tests capture some form of intellectual capacity, they are measuring a form of *previous learning rather than learning ability*[11]. Valid assessments of learning potentials instead would require embedded or dynamic assessments (Resing et al., 2020; Tzuriel, 2012), i.e, assessments that probe children's learning progress over time. Thus, it is clear that CITO can neither serve as a test of learning potentials nor as an intelligence test: On the one hand, since the test is static and not embedded in the teaching process (cf. Sect. 2.1.2), it cannot be used for dynamic assessments. On the other hand, to test IQ, multiple-choice questions would need to be formulated in an unambiguous way, and answer choices would need to be unambiguous when the tested skill is applied. Yet as illustrated by the examples in Appendix A.3, CITO questions and answers are far from unambiguous in all these respects. CITO questions furthermore heavily rely on *language vocabulary*, in both text comprehension as well as math skills ('verhaaltjessommen'), as well as on *prior knowledge*.

Could CITO then at least serve as a valid test of skills and learning progress? It turns out that even this is problematic. CITO tests do not necessarily measure the skills related to learning goals (e.g. in algebra or text comprehension), but often rely on other *contingent* skills[12]. To render a valid skills test, CITO would thus require *special training*, preparing students in terms of those contingent skills and the particular way how a question is posed. However, this training is precisely prevented by the way the test is performed, as explained in the following.

Thesis 1. CITO tests neither measure intelligence nor are they suitable for assessing learning potentials, because they do not comply with principles of question clarity and with dynamic assessment. Furthermore, CITO rely on *contingent skills*. Using CITO tests for valid assessments of skill levels would thus require prior training.

2.1.2. Lack of constructive alignment of tests

Learners need to actively construct knowledge rather than just passively consume information. Starting from our own individual capacities, we learn by incorporating *regular feedback* into our own schemas to adapt them accordingly and to build our own representations (Glasersfeld, 1995). However, decades of *constructivist pedagogy* seem to have led to the mistaken view that such learning would happen autonomously and without control of a teacher (Richardson, 2003), whose role shrinks to a mere 'facilitator' who only offers learning material (Biesta, 2015). Yet, teachers have an essential and very active role to play in this learning process. Students can indeed learn something *from* their teachers, yet this means that teachers need to *teach* (Biesta, 2020), i.e., they need to give instructions and feedback to be able to steer the learning process. Furthermore, learning also needs to proceed *in a particular order*. The reason for the latter lies in another (seemingly forgotten) constructive principle, namely the principle of *methodical order* (Janich, 2001). According to Janich, there is a certain ordering to every cultural method which derives from its constructive practice[13]. This applies to everyday skills, e.g., when you learn to pull on your trousers before your shoes and not vice versa, as well as to educational skills. In mathematics (Glasersfeld, 2006), it is e.g. important to first learn how to measure the area of a rectangle (as a product a * b) before we learn to measure the area of a triangle (1/2 * a * b), simply because the latter can be explained by fitting exactly two congruent triangles into every rectangle. An in-depth understanding of methods is not possible when this methodical order is violated.

For these reasons, the only valuable way of measuring learning is to make the assessment part of the learning process (Biggs, 1996). Valid assessments of the level of acquired skills can only be done when the assessment is *aligned* with the learning process. This means that those and only those capabilities should be tested that were explained to and practiced by a student in the required methodological

[9] https://nl.wikipedia.org/wiki/Centraal_Instituut_voor_Toetsontwikkeling.

[10] https://www.kidsweek.nl/voor-ouders/heeft-oefenen-voor-de-cito-toets-zin.

[11] https://www.leidenpsychologyblog.nl/articles/is-intelligence-testing-in-school-the-smart-thing-to-do.

[12] With contingent skills, we mean skills that are not required for the tested learning goals. For example, when a math test presupposes language skills not required for the tested math skills.

[13] Note that the principle of methodical order does not imply that learning must be linear. In fact it is often cyclical, e.g., according to Kolb's experiential learning cycle (Kolb and Kolb, 2018).

order in the period before the test. In a nutshell, the order of testing needs to follow the order of instruction and practicing, otherwise we risk to merely measure differences in habituation with certain types of questions. We call this fundamental principle *constructive alignment*, following Biggs (1996).

Yet, the reality of testing in Dutch primary schools is a continuous violation of the constructive alignment principle. CITO tests are standardized contentwise and held at regular half-year intervals. However, this is done in a way that makes it deliberately hard to align with the learning process. Tested skills are taken from a mix of subject areas learned over longer periods, intermixing e.g. algebra and geometry with measurement units in unforeseeable ways[14]. This makes it hard to prepare for a CITO test or to use it as an assessment method for testing a particular acquired skill level. There is thus a high chance that students come across questions they have not fully grasped because topics do not necessarily follow a methodical order. For example, tests in primary school sometimes contain geometry tasks that require pupils to calculate areas of triangles or volumes with triangular shapes. Yet, it is clear that a methodical order would require understanding the formula (1/2 * a * b) first. The latter, however, is only fully understandable when it is motivated appropriately in geometry lessons in secondary school. Similarly, primary school tests may contain questions requiring children to simplify fractions without having fully grasped fractional algebra. This is because the function of the CITO is not to give feedback to students or teachers, but is seen as a way to divide pupils into different educational streams, as explained in the following.

Thesis 2. CITO is not a valid system for testing the skills of students because it is not *aligned with the learning process*. Different subject areas that are learned at different times are mixed in a test. Ambiguous questions would require practising, however practising is not intended. More generally, CITO fails as a test for skills as well as potentials because it has never been intended as a feedback instrument.

2.1.3. Lack of absolute reference norms

Grading systems are necessarily based on norms that let us evaluate performances as unacceptable, acceptable, good or very good. Depending on how we arrive at such norms, we can distinguish *absolute, individual* and *social reference norms* (Elliot and McGregor, 2001; Ingenkamp, 1977; Rheinberg, 2001a). Social reference norms measure a test result relative to the distribution of results of some peer group. For example, we may say the result of some student is good if it is better than 80% of the results of all students of the same age. A similar social reference norm is used in all CITO grading systems in the Netherlands, defined as quantiles of points reached in a test by all students of a given grade[15] (Fig. 3a).

In contrast, individual reference norms measure achievements relative to the individual learning history. For example, a teacher might measure the increase of words a child knows in a foreign language. Absolute reference norms, instead, measure a test result using minimal standards defined for concrete learning goals[16]. For example, to obtain a pilot license, it is not sufficient to know whether a candidate belongs to the best within a group. Instead, it is important to make sure that the candidate has certain minimal theoretical and practical skills for flying a plane. Only absolute norms can assure that successful students are capable of successfully performing a particular practice, of 'jumping over the ditch' (Ingenkamp, 1977; Rheinberg, 2001a). Absolute reference norms are therefore the only norms that can provide a standard comparable across groups. Unfortunately, they are largely lacking in standard grading systems across Dutch schools.

Using social reference norms leads to serious inconsistencies and misconceptions of learning potentials (Rheinberg, 2001b), as well as to unfair distributions of educational opportunities. It leads to inconsistency because the same skill level might be evaluated as 'good' or 'insufficient' depending on the choice of the reference peer group. Since CITO scores are dependent on the result distribution of comparable age groups of a certain year, the same performance might be judged differently in different years. In the most extreme case, if all students reached a similar CITO result, still only 20% would score high, *missing out on 80% of comparably capable students*. In addition, social reference norms ignore learning processes that are achieved by the entire group, as well as by individuals relative to their learning history (Rheinberg, 2001b). It is well known that social reference norms therefore can demotivate students in their learning process (Dickhäuser et al., 2017; Rheinberg, 2001a). To illustrate why, suppose a student has a backlog ('achterstand') in text comprehension skills due to an immigration biography. This is not an exotic case, since roughly 25% of all inhabitants of the Netherlands have a migration background. Since CITO presupposes a significant mastery of language even to understand math tasks, it is considerably harder for such a student to achieve levels comparable to mother tongue peers. When using social reference norms, such students will likely be outside of the best 20%, even if their skills would be sufficient for mastering vwo. Social reference norms manifest these differences by devaluing good performances, and thus help path the way towards an elitist educational system.

Thesis 3. The CITO grading system is demotivating and elitist because it measures success entirely in terms of social reference norms, not in absolute reference norms. This makes the test invalid as a means for measuring skill levels and exploiting learning potentials.

2.1.4. Stereotypical qualitative criteria for learning potentials

In addition to tests, Dutch teachers also use qualitative criteria for assessing the suitability of a student for different educational

[14] Sometimes it is argued this delay in CITO is intentional because it would test a student's memory. However, this is not a convincing argument since valid memory tests require *specific memorizing instructions* which are lacking in the case of CITO.

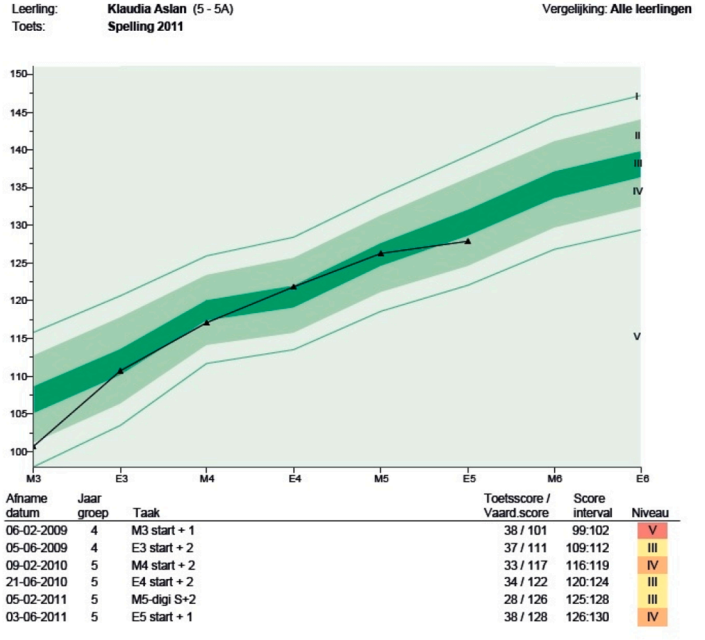
[15] The sample on which these quantiles are computed can be up to 6 years old.

[16] It is sometimes argued that LVS systems would be based on absolute reference norms because test questions are made with 'referentieniveaus' in mind. However, this is a misconception: The norm for measuring CITO grades is based on social references.

I–V		A–E	
20% hoogst scorende leerlingen	I 20%	A 25%	25% hoogst scorende leerlingen
20% boven het landelijk gemiddelde	II 20%	B 25%	25% ruim boven tot net boven het landelijk gemiddelde
20% landelijk gemiddelde	III 20%	C 25%	25% net tot ruim onder het landelijk gemiddelde
20% onder het landelijk gemiddelde	IV 20%	D 15%	15% ruim onder het landelijk gemiddelde
20% laagst scorende leerlingen	V 20%	E 10%	10% laagst scorende leerlingen

(a) CITO scores, defined as quantiles over the nationwide distribution of CITO points reached by students of a given grade.

Figuur 1 Voorbeeld van een leerlingrapport



(b) How Cito results are reported in school reports. Y axis denotes CITO points, green areas denote quantiles defining scores, and the black line denotes a student's performance over time.

Fig. 3. The CITO grading system.

levels. These 'soft' criteria are not officially recommended by the inspectorate of education[17] because they lack any scientific methodological basis. Still, they are used regularly in different places, e.g., in terms of 'leerlingprofielen'[18] or as part of student reports ('Onderwijskundig rapport') by primary schools. Though it is currently unknown how widely spread such criteria are, education professionals told us in personal exchange that they are considered a proven tool of practice, especially in situations of doubt over the capabilities of a student. One example is shown in Fig. 4. As can be seen, the criteria cover study skills ('vaardigheden'), social-emotional ('sociaal emotioneel') characteristics as well as working attitude ('werkhouding'). Here is a translated summary for the level of vwo/havo as opposed to the level vmbo-b (highlighting by the authors of this article):

- Study skills: vwo/havo students are supposed to work on *abstract tasks*, can apply *abstract concepts*, can reflect based on their own judgement, and can draw conclusions. Students on the vmbo-b level are supposed to learn only in *concrete terms* and *in the 'now'*.
- Social-emotional characteristics: vwo/havo students are supposed to use others' feedback and can criticize other people's reactions. They are also *less influenced by others* and *less dependent on confirmation by teachers*. Students on the vmbo-b level are expected to see school primarily as a *social event*.
- Working attitude: vwo/havo students *can concentrate on their work for at least 30 minutes*, can *make their own plans*, are *assertive* and can *lead group tasks*. Students on the vmbo-b level can *concentrate only for 10 minutes* and need *guidance from teachers*.

In this table, we can see that CITO and other test results are treated equivalently with intelligence test (IQ) scores. From a scientific standpoint, this mapping is highly questionable for the reasons outlined above.

Furthermore, there are numerous reasons to doubt the validity of these qualitative criteria for the purpose of assessing learning potentials. For instance, regarding study skills, it is well known in Psychology that 'being absorbed by the now', denoted by the technical term *flow*, is a very important concept for learning as such, regardless of the educational level (Csikszentmihalyi, 1990; Engeser and Rheinberg, 2008). It is known that literally everyone experiences a flow when learning 'in-depth', not only people with low learning potential. For this reason, it is questionable to take such a characteristic as a criterion for differentiating learning potentials. Furthermore, the idea to differentiate low and high learning potentials based on distinguishing *concrete* versus *abstract* or *theoretical* concepts reflects another misconception, one that has become famous as a so-called *category error* by Ryle (1949): Even in the most theoretical sciences, like mathematics, knowledge is not gained abstractly. Knowledge builds on know-how, i.e., the disposition to operate on concrete material. This means that we can learn theoretical concepts only by applying them in concrete practices[19] (Janich, 2001; 2015). Therefore the distinction between practice and theory is *essentially unsuitable for differentiating* educational skill levels.

The idea of students on vwo/havo levels supposedly being more *asocial* in their behaviour, i.e., less dependent on others and their teacher's guidance, reveals another questionable stereotype underlying this assessment scheme. In-depth understanding of a concept always requires not only practice but also critical guidance and interaction with teachers and peers (Janich, 2001). The essential role of social relations for education is highlighted by numerous researchers, e.g., by the social dimension of Biesta's triangle (Biesta, 2020), by the extensive empirical investigations of the pediatrician Remo Lago (Lago and Beglinger, 2009), as well as by learning theorist Etienne Wenger (Wenger, 1999). Since even the most intelligent students need guidance and feedback in their learning, the idea of the 'autistic gifted' student is largely a myth. Or as Gooch (2019) puts it:

> Since the knowledge gained comes primarily through interrogation of and by others, education is *relational*, depending on personal interaction between teacher and student.

Even the qualitative criteria about *working attitudes* on different educational levels are hard to defend. Though the ability to concentrate and the ability to self-structure and to make plans form surely an important requisite for mastering a university level, they are *not a sign of lack of intelligence or learning capacity*. This can be seen from the fact that concentration and self-organization is not only a problem for slow learners, it is also one for intelligent and gifted students or students with Attention Deficit Hyperactivity Disorder (ADHD). Around 30 per cent of the gifted children in the Netherlands ('hoogbegaafdheid') (with an IQ higher than 130) end up on an education level below havo, i.e., they clearly *underachieve*[20]. Furthermore, *underachievement* is not only a problem for the gifted. There is a significant amount of children who do not underperform because they lack learning potential, but because they lack self-confidence, have difficulties concentrating and organizing themselves (Dittrich, 2014).

Thesis 4. Qualitative criteria based on student characteristics ('kenmerken', 'leerlingprofielen'), including concentration levels, social dependence and working attitudes, are invalid as a means to assess a student's learning potential.

2.2. Teaching process and teaching paradigms

Addressing the problems of assessment mentioned in the previous section cannot be done by merely redesigning tests, or by improving assessment norms or criteria. It needs to be accompanied by a redesign of the educational process in class and a change of

[17] https://www.onderwijsinspectie.nl/onderwerpen/overgang/welke-factoren-meewegen.
[18] https://swvbepo.nl/.
[19] This can also be a mental practice, e.g. the practice of manipulating mathematical expressions.
[20] https://centrumregenboogkind.nl/hoogbegaafdheid/.

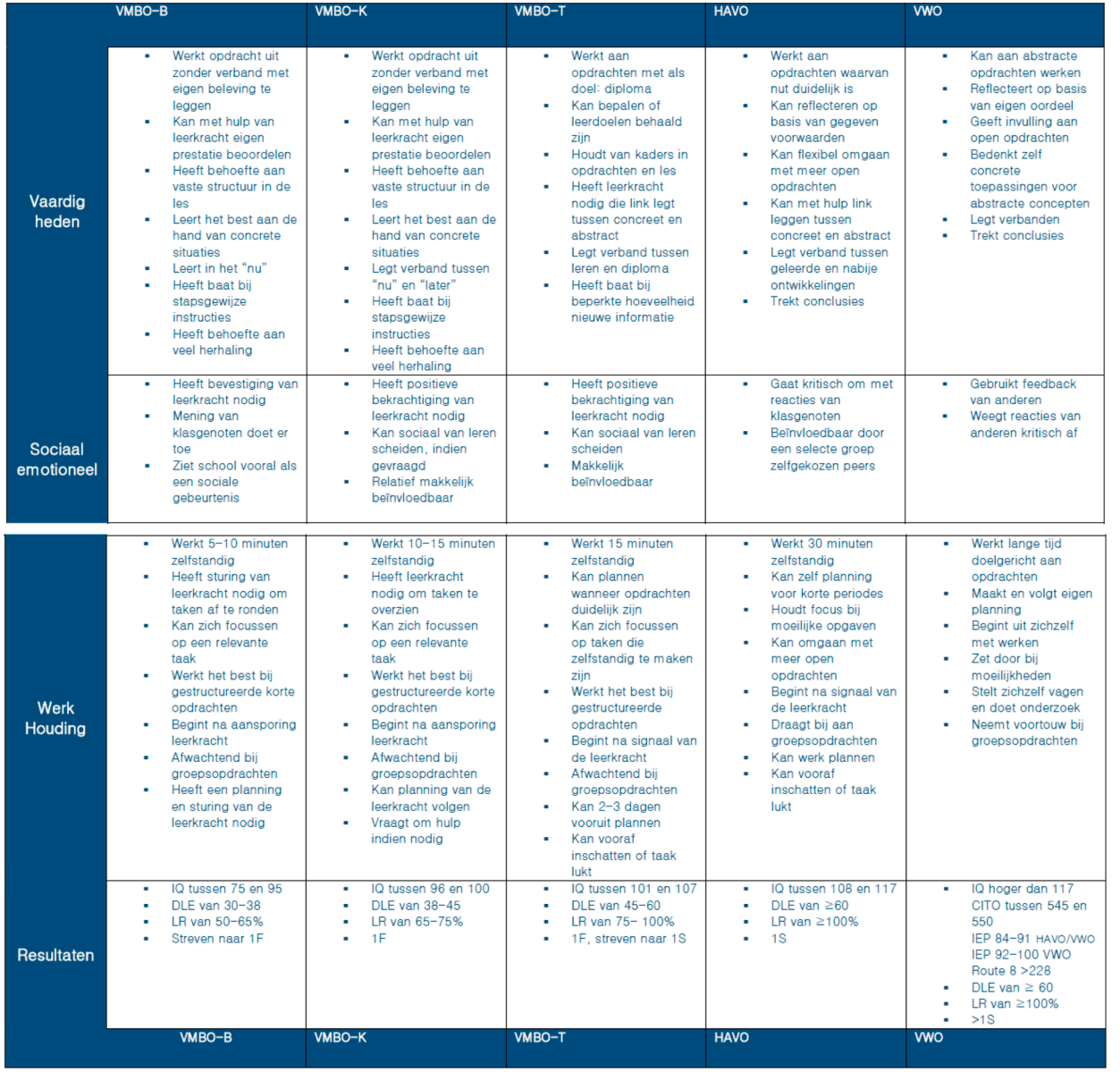


Fig. 4. An example of a student profile ('leerlingkenmerken') for assessing the suitability of Dutch students for educational levels (Source: https://www.groei-onderwijsadvies.nl/).

the underlying teaching culture (Heij, 2021). In the following, we discuss this culture by contrasting different *paradigms*[21].

2.2.1. Test vs. learning goal constrained teaching

For example, to make sure that tests are constructively aligned with the teaching process and to make use of absolute reference norms, it would be required to standardize concrete learning goals and to translate them into teaching and testing practices in class[22]. However, if we take a closer look at the national attainment targets and reference levels (cf. Appendix A.1), we see that attainment targets reflect only general topic areas[23], rather than concrete, measure-able goals. Furthermore, though reference levels are concrete enough in this sense[24], they cover only very restricted areas of learning (Heij, 2021), namely only applied algebra and language skills. Thus concrete standardized goals are missing for essential forms of knowledge taught in school, such as Geometry, Chemistry, Physics, Biology, oral and argumentation skills (Osborne et al., 2004), civil and political education, History, Geography, artistic abilities, etc. In addition, though standardized tests were designed with reference levels in mind, teachers reuse standard tests without being made aware of their underling targets. Reference levels thus tend to *remain hidden in test methods*, and tests fail to cover the breadth of the actual teaching practice. The relevant tests (the ones used for selection) are therefore usually not designed by teachers, they are predetermined by LVS standards[25]. Thus the teacher is neither expected to design class according to such tests, nor is he or she expected to design tests corresponding to learning goals. We call this paradigm *test constrained* teaching, as opposed to *learning goal constrained* teaching. In the former, a teacher's responsibility ends with a test, whose design is beyond his or her own responsibility. In the latter, a teacher's responsibility includes the design of tests as a means of teaching, and therefore ends only with predefined learning goals. To allow learning goal constrained teaching, *binding* goals need to be tested by teachers *in their own responsibility*, and *in their own timing*. This is because teachers are the only ones who oversee the learning process of children in a class.

Thesis 5. Teachers lack explicit, concrete, binding and testable learning goals for essential areas of knowledge. The concrete goals of reference levels are limited in breadth and hidden in test methods. This prevents both alignment of skill tests with teaching as well as embedding assessments of learning potentials, and thus sustains invalid assessments on all levels.

2.2.2. Gate keeper role vs. guarantee of a minimal level of education

The role of a teacher in Dutch schools resembles that of a *gate keeper* for educational levels. The gate is opened once there is *proof of a socially exceptional performance*, and in all other cases, the gate needs to stay closed. Without doubt there are many teachers who do their best to support students in their ambitions, offering learning material and explanations to this end. But teachers are usually not made responsible for the achievement of minimal learning goals, and thus passing the gate lies entirely within the responsibility of the child. In consequence, this responsibility is not taken over by anyone, except by those parents who understand this situation and can afford private tutoring (Elffers, 2018).

Thesis 6. The teacher's responsibility involves performance oriented gate keeping which conflicts with the role of granting a minimal level of education corresponding to the learning potential. This undermines the exploitation of learning potentials.

This paradigm shows in various teaching practices. One example concerns the responsibility of preparing students for a test. It e.g. is hardly understandable why university students frequently obtain pre-exams to prepare for an exam, as well as clear instructions on content limits for preparing themselves (McKeachie and Svinicki, 2013), but pupils in CITO tests in general do not. Preparations for a test are thus dependent on a teacher, a school type or even the money or time parents spend on private tutoring (Elffers, 2018). Also, teachers are usually not made responsible for *the behavior of pupils in class*. Yet, the control of the learning environment in school is essential to make sure all pupils have a chance of following (Schultz, 2009), not only the ones who are taught the content at home after school or in private tutorials. Pupils who cannot follow the lesson because of social distractions in classroom ('class room climate') are known to underachieve (Dittrich, 2014)[26], yet these students can learn to concentrate in the right environment. To be effective, exercise and training needs to happen under controlled conditions. Correspondingly, giving homework improves pupil achievements only in families of upper socio-economic scales, whose parents are able to exercise such control (Rønning, 2011).

2.3. School policy and social paradigms

Our methodical critique so far has focused on the assessment and teaching processes, but not on the political and societal conditions in which they are embedded.

[21] A paradigm is a predominant thought pattern of a culture.

[22] Establishing test validity becomes a challenge if such goals are not standardized, e.g., on an international level (Wolf, 1998).

[23] See e.g. Besluit kerndoelen onderbouw VO of 2006, https://www.rijksoverheid.nl/documenten/besluiten/2010/09/17/kerndoelen-onderbouw-voortgezet-onderwijs.

[24] For example, in 'rekenen', 'referentieniveaus' across levels 1F/S to 3F/S include the handling of numbers, ratios, measurements and tables/diagrams. For primary schools, similar concretizations are formulated in 'Inhoudslijnen' (TULE), see https://www.slo.nl/sectoren/po/inhoudslijnen-po.

[25] 'Methodentoetsen' are foreseen as a kind of formative testing designed by schools, however, they are usually regarded as less relevant for selection purposes.

[26] The Munich Model of Giftedness and Talent (Ziegler and Heller, 2000) includes classroom climate and quality of instruction as important moderator variables for explaining underachievement.

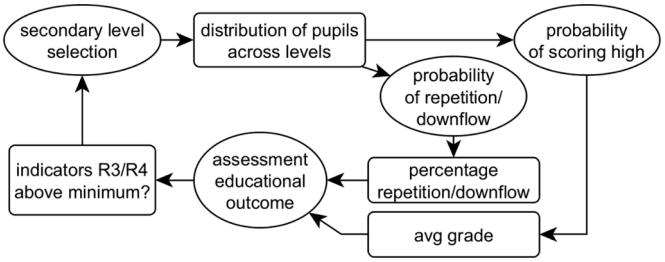


Fig. 5. Unintended feedback of educational outcome assessments R3/R4 on the selection into levels in secondary school. For a school, distributing pupils into lower levels is advantageous because it will increase the probability of scoring high in R3 and R4.

2.3.1. Unintended feedback in indicator driven school evaluation schemes

The Dutch educational policy is driven by numeric indicators. For example, to judge the quality of secondary schools in the Netherlands, the Dutch inspectorate of education has established statistical norms called the *educational outcome model* ('Onderwijsresultatenmodel')[27]. This model[28] consists of 4 different indicators of educational success for a secondary school:

1. *R1* ('onderwijspositie') measures the difference between upflow (the percentage of pupils streaming from lower levels to higher levels) and downflow (the percentage of pupils streaming from higher levels to lower levels) in year three of secondary school with respect to the level of the recommendation in primary school. The more positive the difference, the better.
2. *R2* ('onderbouwsnelheid') measures the percentage of pupils that reach year 3 without repetition. The higher, the better.
3. *R3* ('bovenbouwsucces') is a combination of both factors, the upflow/downflow difference and percentage without repetition, measured from year 3 for each level for the remainder of the school years (after year 3).
4. *R4* ('gemiddeld CE') is the average score of the central exam over all exams written in the final year.

On first look, these criteria all seem to make perfect sense. The Dutch society indeed wants schools to increase upflow and prevent downflow, to prevent pupils from repeating school levels, and to make pupils score as high as possible in the final exams. Schools who 'succeed' in realizing this should be rewarded, because they obviously are doing something right, or so it seems?

However, on a closer look, there are severe problems with this scheme. For one, the different indicators are dependent on results of standardized tests which were designed for pupil selection purposes, not for evaluating schools, and thus *cannot be used as a valid test of school performance at the same time* (Heij, 2021). Thus, there exists a validity conflict on the level of goals when using a test for both purposes. Furthermore, in the cybernetic system of Dutch school evaluation, the intended effect becomes easily nullified due to *unintended feedback*. To see why, think for a moment how the establishment of such a system will feed back on the behavior of responsible personnel and teachers in a school (cf. Fig. 5). For sure, the responsible persons will try to minimize the chance of falling below the threshold levels defined by the inspectorate of education. For factor R3, e.g., the threshold is currently supposed to lie at 76, 99% (havo) and 80, 58% (vwo). However, the factors that influence the probability of upflow, downflow, repetition, and the distribution of average grades on a school are largely influenced by social preconditions in the school's neighborhood which are beyond the influence of a school. Exactly this effect can be seen by the fact that the inspectorate of education was forced to introduce corresponding *social correction factors* for their threshold levels[29]. For this reason, a school will try to minimize its own risk simply by *being conservative in its assessments* of those students. Without any doubt, distributing children into lower levels in *secondary level selection moments* (e.g. in the 3rd year of secondary school) raises their probability of scoring high and decreases the probability of repetition and downflow, and thus improves the criteria R3 and R4. A similar effect can be observed for R1 and R2, because primary schools tend to strategically anticipate the decision process on secondary level[30]. In consequence, if a school followed a more liberal selection policy at a larger scale, it would risk the lowering of average grades, the rise of repetitions, and a decrease of upflow. Schools therefore tend to become distrustful of the learning potentials of their students. A visible sign of this distrust is the comparably high percentage of children who need to repeat a grade in secondary school to make sure they succeed later on[31]. In 2012, the percentage of secondary school graduates who repeated at least one grade was 45 %, on havo level even 52% of the graduates (Van Vuuren and Van der Wiel, 2015). Two third of these repetitions occurred in secondary school. *So approximately half of all graduates had to repeat a grade*. Compared to other countries in the same year, 27,6 % of all Dutch 15-year-olds had already repeated at least one school grade, which is approximately twice the OECD average (Zapata et al., 2014).

[27] https://www.onderwijsinspectie.nl/onderwerpen/onderwijsresultatenmodel-vo/indicatoren.

[28] Sometimes, such scores are also discussed under the term 'slagingspercentage', however, the latter refers only to the percentage of successful final exams, which appears as one factor in R3.

[29] https://www.onderwijsinspectie.nl/onderwerpen/onderwijsresultatenmodel-vo/indicatoren.

[30] This behavior is recognized under the term 'strategisch gedrag'. In fact, the inspectorate of education has realized this problem and as a consequence stopped the evaluation of teachers based on LVS results (Borghans and Schils, 2015; Muskens and Tholen, 2015).

[31] Though repetitions are also among the indicators, they are recognized as means to improve school performance (Van Vuuren and Van der Wiel, 2015), because they can improve the other indicators (downflow as well as average grades).

Table 1
Amount of graduates per year with scientific, applied and vocational diplomas in the Netherlands from 2010 till 2019. Percentages of these three levels are given from total. For 3rd level education: wo includes university-level bachelor, master, doctoraal and beroepsdiploma. Applied university diplomas (hbo) and vocational degrees (mbo). Source: CBS Statline, 'Gediplomeerden en afgestudeerden; onderwijssoort, vanaf 1900', https://opendata.cbs.nl/statline/#/CBS/nl/dataset/80384ned/table?dl=59925

2nd level	**2010**	%	**2013**	%	**2016**	%	**2019**	%
VMBO	91,602	55.0	96,332	55.8	103,203	59.8	99,294	52.0
Havo	42,404	25.4	44,560	25.8	47,063	27.3	53,159	27.8
VWO	32,641	19.6	31,816	18.4	34,478	20.0	38,455	20.1
Total	166,647		172,708		184744		190,908	
3rd level	**2010**	%	**2013**	%	**2016**	%	**2019**	%
Mbo	144,947	52.8	152,164	53.5	142,198	49.4	140,226	48.0
Hbo	61,633	22.4	60,913	21.4	68,362	23.7	67,403	23.1
WO	67,961	24.8	71,520	25.1	77,417	26.9	84,390	28.9
Total	274,541		284,597		287,977		292,019	

Thesis 7. The Dutch school evaluation and rewarding scheme in secondary school favors conservative, elitist pupil biographies to increase the probability of scoring high in educational outcome assessments. This conflicts with the assessment of learning potentials. In general, the use of results of selection tests as a means to evaluate school performance makes the scheme invalid in the first place.

2.3.2. Fixed quotas and meritocratic justification of education

During the last century, the educational policy in the Netherlands has lead to fixed educational quotas which became manifest in stable graduation percentages (cf. Table 1): In 2010, only 19.6% of students graduated on vwo level, 25.4% on applied university (havo) level, and the *large majority of students (55%) was educated on the vocational level*. Graduation numbers on the third level repeat this pattern[32]. These percentages mirror the situation of recommendations given in Amsterdam 1967 after de Groot's invention of CITO and have remained unquestioned since then (cf. Appendix A.2).

Though they are an outcome of historical legacies and conservative political decisions (Heij, 2021), quotas are stabilized precisely by the factors that were discussed above, in particular by the predominance of social reference norms.

However, in other European countries, a much larger proportion of students graduate on academic or applied university levels. For example, in Finland, 54% of students graduated in 2019 on the level of upper secondary school, which enables these students to directly study at a university, and 45 % on the vocational level (McFarland et al., 2019). Note that a large part of the latter students still graduate on a level that allows them to visit applied university. In 2020, 35% of all German secondary school students graduated on an academic level ('Gymnasium'), 45% on a medium educational level ('Realschule'), and only 15% on the vocational level ('Hauptschule') (Bildungsberichterstattung, 2020). Again, a significant part of students on the 'Realschule' are able to later graduate on a level ('Fachhochschulreife') that allows them to study at applied university. These numbers are the result of an enormous educational expansion in Germany since 1960. At that time, only 6% of secondary school students graduated on the university level and more than 50% on the vocational level[33]. In the past, such expansions have considerably decreased the difference in skill levels between social groups (Keeves et al., 1991). It seems that a similar educational expansion (in terms of quotas) has never happened in the Netherlands.

Since Dutch pupils are not less capable than their Finnish or German fellows, it must be the case that a significant part of the students that do not attain academic or applied university levels when graduating from secondary school would nevertheless be capable of doing so. Part of these students on havo and vmbo levels are actually trying to make their way up, and some succeed. Research has shown that only 69% of all students who did not repeat a grade were still within their recommended level in the 4th year of secondary school. Note, this also means that roughly a third of these students are obviously subject to invalid assessments. Approximately 17.7 % moved to lower levels, and 13,6% of students moved up to higher levels (Van Rooijen et al., 2017).

The policy of fixed quotas entails that education becomes a kind of *zero-sum game*, where the amount of students with high learning potential which exceed these quotas cannot get access to higher education. This is only justifiable using a radical *meritocratic principle*, namely the idea that, *regardless of their potential*, education needs to be *achieved by, and not granted to*, students (Sandel, 2020). However, this principle not only contradicts the idea of education as a human right, it also reflects a dangerous societal trend. Drifting towards a meritocratic ordering, the Dutch society is generating new social classes of 'losers' and 'winners' (van Pinxteren and de Beer, 2016). A corresponding elitist attitude can be found in some Dutch schools (Boterman, 2013; Merry, 2019), including the 'state-sponsored elite' education at the Dutch 'gymnasium' (Merry and Boterman, 2020). This is dangerous for several reasons, cf. Sandel (2020): For one, when education becomes an elite club to which many capable students will have no access, these students get frustrated to the point where they lose self confidence and therefore are likely to underachieve (Leeuw and Lecina, 2016). And second, such a system punishes autonomous/critical and rewards streamlined personality traits on all levels, as was found in a representative

[32] There is an increasing trend on the university level, however this is probably due to an increase of foreign students.
[33] https://www.fr.de/wissen/schulabschluesse-deutschland-11730979.html.

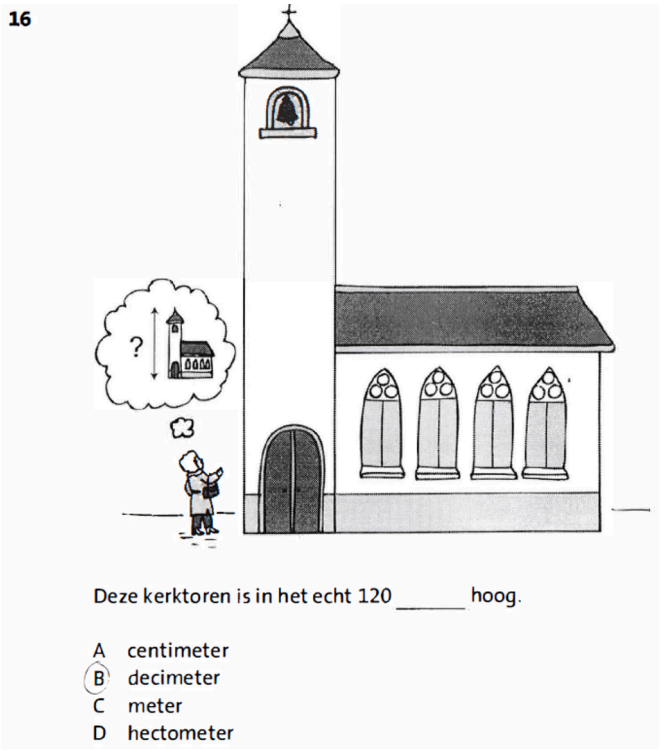


Fig. A1. Task from a CITO test in basic algebra ('rekenen'), group 8.

Dutch study (Lubbers et al., 2010)[34]. This undermines the very principles of an open society (Popper, 1971). To avert such threats, education should not be seen as a matter of private fortune, but rather as a common good.

Thesis 8. The privilege of academic education in the Netherlands is reserved for a fixed percentage of students due to arbitrary quotas which have remained unquestioned for decades. The underlying meritocratic justification is radical: it effectively denies a significant part of society access to educational pathways and prevents effective exploitation of learning potentials.

3. Conclusion and outlook

In this article, we have investigated possible reasons for the increasing educational inequality and the general decrease of skill levels in the Netherlands. The article was written in the form of a critique, which is a way to reflect on the validity of methods of judgment. We explained why such judgments may go wrong on various levels of the Dutch school system and why this conflicts with the central goal to provide equal access to education guaranteed by law.

Our arguments highlight the various forms of misjudgment that keep students from exploiting their learning potentials. Since such reasons are not understood well enough, professionals can still assume that, though being far from perfect, the current system would 'work' in principle. And correspondingly, many Dutch conclude it would be their own 'fault' or 'fate' if their children do not succeed in the school system. Our arguments also shutter further assumptions in the public debate.

For example, it is often argued that since CITO is standardized, it would provide an 'objective' way of selecting students into levels, as opposed to the 'subjective' assessment of a teacher. In particular, it is argued that teachers could be more biased than CITO against students with a migration background. However, as we have argued above, there are many reasons to doubt the validity of CITO both as a way to judge student skills and learning potentials, as well as way to select students on this basis. Since an invalid test is always biased, no matter how much it has been standardized, we would be trading bias for bias. This argument is therefore not convincing. Furthermore, assuming that teachers' assessments are biased just because they are fallible humans or because they have diverse backgrounds reveals a dangerous prejudice against teachers and their profession. It not only runs against the very idea of the Dutch freedom of education, it also shows a *naive conception of objectivity* and a *naive trust in quantification*. Instead, it should be clear from our discussion that 'objective' (i.e., trans-subjective) judgments are just a consequence of validity (Janich, 2001), which needs to be defined relative to a purpose. Once a valid assessment is in place, it will also be unbiased for this purpose, regardless of who designed it

[34] "Principally, our results showed that critical approaches to learning, often applied by autonomous students, are not rewarded in Dutch secondary education. Checking other sources and forming one's own opinion appear to affect grades negatively" (Lubbers et al., 2010).

or who is assessed, and regardless of whether it is a standardized, quantitative test or not. Stereotypical qualitative assessment schemes, as currently in use, do not help because they lack any methodical validity.

From a research perspective, further empirical research is needed to substantiate our theses and to gather more evidence for the observations that have been made. For example, it is currently unknown how widely distributed stereotypical qualitative assessment schemes are, and *to what extent* quantitative tests such as CITO are dependent on prior contingent skills. An important question is how concrete learning goals need to be in order to serve goal constrained teaching. From a didactic viewpoint, educational researchers should investigate how tests can be aligned with these goals and the teaching process, and how dynamic assessments would alter the estimation of potentials. Furthermore, the detrimental effects of indicator driven school evaluations and fixed quotas on the biographies of children should be further investigated by comparative cohort studies. Finally, a far reaching consequence for research is that statistical studies (cf. Dronkers and van de Werfhorst (2016)) which use data about achievements within CITO-like systems *cannot be used to draw conclusions about skill levels and learning potentials* of Dutch students.

From a political perspective, we have argued in this article that valid assessments require, on the one hand, the definition of concrete and testable learning goals, and on the other hand, the embedding of assessments into goal constrained teaching. The first would require the Dutch society to standardize concrete goals across all areas of knowledge and, in turn, to let go of standardized tests. Thus, not the testing or teaching needs to be standardized, but the learning goals. Testing needs to be given into the hands of teachers as part of their responsibility for achieving these goals. It might be argued that both teachers and the society are overwhelmed with this task. However, the Netherlands is the only country in Europe that selects secondary school students into educational levels based on a standardized national test (Naayer et al., 2016). In other European countries, the selection lies in the hands of teachers and is based on repetitive assessments of whether pupils have reached minimum learning goals required for passing on to the next grade. Furthermore, since the selection advice given at the end of primary school is usually non-binding (Naayer et al., 2016), pupils can try out academic level education to disprove invalid assessments of their learning potential. To realize this in the Netherlands might require the strengthening of the academic education of teachers and a raise of their salary. On the level of school policy, first, it is required to rethink how schools are evaluated, because the current scheme is not only invalid as a means to assess a school's quality, but also minimizes the chances of growth of its students due to unintended feedback which reinforces conservative assessments. And second, if evaluations are redesigned by dropping social reference norms and adopting absolute reference norms, fixed quotas need to be abolished, because the number of students on a certain educational path can no longer be determined a-priori. It is an outcome of valid assessments. Finally, such reforms need to be embedded into a public discourse over humanistic education in the Netherlands, a country with one of the oldest liberal traditions in the world.

Acknowledgements

We are very grateful to Martin Uunk, Greetje Timmermans, Karen Heij and Frank Ostermann for their valuable suggestions regarding earlier versions of the manuscript. We thank Jessica Hilhorst, Annette Ter Haar, Claire Boonstra and all other members of the BOT group for their continuous support. We also would like to thank the anonymous reviewers for their helpful suggestions. Finally, we thank all members of Studio Moio, in particular Eva de Leeuw, Mitchell Kort, Laila Aissi, Thijn van der Geest and Rashad Mashaal, for discussing and translating our article.

Appendix A. The Dutch school system in a nutshell

In this section, we summarize the most relevant characteristics of the school system and shortly illustrate the most important standardized test.

A1. Overview

The variety of Dutch school forms, including public-authority schools, Catholic, Protestant, Montessori, Waldorf, Dalton, and Jenaplan schools, is due to the principle of *freedom of education* ('Vrijheid van onderwijs') which has its roots in the division of the Dutch society into cultural and religious *pillars* ('verzuiling') (Spiecker and Steutel, 2001). In consequence, schools have been free to choose their teaching methods as well as learning contents, while being *publicly funded* for both public-authority and private schools. Main goals are formulated in nationwide guidelines which are binding for all schools and which focus mainly on algebra and text comprehension skills, as laid down in current national laws (Secondary Education Act[35] and its *attainment targets* ('kerndoelen')[36] and *reference levels* ('referentieniveaus')[37]. To assure students meet the performance requirements specified in these guidelines, *state-wide standardized tests* are held regularly as part of a *student tracking system* ('leerlingvolgsysteem' (LVS)), starting in primary school. These tests shall assure that minimal reference levels are reached across the diverse school landscape. The most frequently used LVS is the CITO system developed by the Centraal instituut voor toetsontwikkeling[38].

Between the ages of four and twelve, children attend *primary school* ('basisschool') running through eight grades ('group 1' through

[35] https://wetten.overheid.nl/BWBR0002399.
[36] https://www.rijksoverheid.nl/documenten/besluiten/2010/09/17/kerndoelen-onderbouw-voortgezet-onderwijs.
[37] https://www.rijksoverheid.nl/onderwerpen/taal-en-rekenen/referentiekader-taal-en-rekenen.
[38] https://nl.wikipedia.org/wiki/Centraal_Instituut_voor_Toetsontwikkeling.

'group 8'). School attendance is mandatory and starts in group 2 at age five. Starting from group 3, twice a year all pupils are tested based on standardized LVS tests around January and in the beginning of June. In group 8 all schools are required to do a final aptitude test called the 'eindtoets basisonderwijs'. Since 2015, teachers have the final authority to decide about school recommendation, yet the 'eindtoets' together with previous LVS test results as well as qualitative assessments are the most relevant criteria used in this decision. The result is a *binding recommendation* which constrains accessibility to the different educational levels of secondary school.

Secondary education, which begins at the age of 12 and is compulsory until the age of 18, is offered at several levels. Vocational training (vmbo) is offered on 4 different levels. Non-vocational education is offered on two levels: havo (five years) and vwo (six years). The havo diploma is the minimum requirement for admission to hbo (university of applied sciences). The vwo diploma grants access to academic universities. There are various ways how students with a lower recommendation can still make it up to university. For example, one can get into university also after successfully completing the first year ('propaedeuse') of hbo.

A2. The evolution of the CITO testing system

In the first half of the 20th century, the Netherlands was searching for a way to prevent pupils from dropping out of school. The idea was to select pupils into educational levels *in a more systematic way* according to their merits and according to society's needs (Heij, 2021). The need for a standardized method of selection was answered by A.D. de Groot's invention of CITO in 1966. Multiple choice questions allowed to standardize and automate the grading (Groot and van Naerssen, 1973). However, based on which learning goals should a decision on educational levels be made (De Groot, 1965), in particular when the society was divided into cultural and religious *pillars* ('verzuiling')? In Amsterdam in 1967, de Groot was involved in setting up a precursor of CITO, the 'Amsterdamse schooltoets'. At the time, approximately 20% of primary school recommendations in Amsterdam lead to an academic secondary level (Heij, 2021, ch. 4.8). de Groot found a correlation between his test scores and these recommendations, and on this basis concluded that his tests would be 'valid' (Heij, 2021, p.140). Thenceforth, the situation in Amsterdam was taken as a *norm for future selection*. This had also the advantage that the limited educational resources (e.g. the limited amount of places available on a level) could be allocated in an efficient manner[39].

A3. Standardized testing with CITO

CITO tests contain usually 3 text comprehension tests about 4-7 texts each, 3 math tests with 32 questions each mostly based on stories ('verhaaltjessommen'), and two spelling tests. In addition, kids read aloud to a teacher while stopping the time and checking for reading mistakes. The questions of a CITO test are *closed ended questions* with *single select options*. This means that only a single one of the offered answers are supposed to be true. The two examples in the following illustrate the difficulty of interpreting CITO tasks and answers unambiguously.

Fig. A1 asks the pupils to assess the metric unit of a given height value (120) for a church tower. The expected answer is 'decimeter', meaning that the tower is supposed to be only 12 meters high, and not 'meter', which would lead to a height of 120 meters. This answer can be guessed by recognizing that the woman is less than 2 meters high and can be stapled six times to measure the tower's height. However, this approach assumes that the depicted drawing is true to scale, which remains unclear from the task description. If this assumption is not made, then a plausible height can also be guessed based on one's experience. Churches in other European countries are often higher than 120 meters. For example, the Cologne Cathedral in Germany is 157 meters, and Notre Dame in Paris is 130 meters high. Depending on the experience of a child, it can be perfectly plausible to assume that the correct answer is 'meter', and not 'decimeter'.

In the following example, pupils need to read a text about the painter Rembrandt (excerpt) and the reasons for his fame:

> The painter Rembrandt van Rijn lived from 1606 to 1669. At that time, the Golden Age, the Netherlands was doing very well. The wealth of the population ensured that the arts also prospered. There were almost 150 painters in Amsterdam alone. People wanted a nice painting on the wall and had the money to buy it. Yet there are only a few of those 150 painters that we still know today, while Rembrandt is world famous today. How come his paintings are sold for millions?... [translation by the authors]

After reading this text, pupils need to answer several single option multiple-choice questions (one reproduced here with our own translations):

Opgave 1. What does the writer want to say with the sentence: *there were almost 150 painters in Amsterdam alone?*

1. That the seventeenth century was a Golden Age
2. That the population was rich
3. That Amsterdam was the capital of art
4. That art was doing well

In this first answer set, 3 of 4 answers are statements truthfully made in the text (1,2, and 4). This means that from the viewpoint of the text content, none of these answers can be considered wrong. For this reason, understanding the content of the text does not help at

[39] Heij e.g. reports about a practice of selection in Curaçao where the percentage was explicitly based on the number of available places in havo/vwo, see (Heij, 2021, p.81).

all in answering this question. Still, the most likely (and expected) answer is the last one, as it is the statement immediately preceding the sentence cited in the question: 150 artists being a sign that art was doing well. However, it can hardly be argued that 150 artists are not equally well a sign for the richness of the population, as well as for the Golden Age. Furthermore, the former fact can rightfully be considered also a sign for Amsterdam as the capital of art during the Golden Age. This ambiguity can only be resolved by measuring syntactical distances between the sentences. *In summary, note that answering both questions asks for special training* because it requires skills that have nothing to do with comprehending a text or understanding an algebra task.

References

Baltussen, M.e.a. (2019). De staat van het onderwijs 2017/2018. *Technical Report*. Utrecht: Inspectie van het Onderwijs.

Biesta, G. (2015). *Beautiful risk of education*. Routledge.

Biesta, G. (2020). Risking ourselves in education: Qualification, socialization, and subjectification revisited. *Educational Theory, 70*(1), 89–104.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364.

Bildungsberichterstattung, A. (2020). Bildung in Deutschland. Ein indikatorengestuetzter Bericht mit einer Analyse zu Bildung in einer digitalisierten Welt. *Technical Report*. wbv Media.

Boonstra, C., de Graaf Bierbrauwer, C., & Carstens, N. (2020). *Het onderwijsvragenboek: waarom doen we de dingen zoals we ze doen?* Amsterdam University Press.

Borghans, L., & Schils, T. (2015). De invloed van opbrengstindicatoren op het functioneren van scholen in het primair onderwijs. *Technical Report*. Onderwijsinspectie, Universiteit Maastricht.

Boterman, W. R. (2013). Dealing with diversity: middle-class family households and the issue of blackand white schools in amsterdam. *Urban Studies, 50*(6), 1130–1147.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.

De Groot, A. (1965). Vijven en zessen.

Dickhäuser, O., Janke, S., Praetorius, A.-K., & Dresel, M. (2017). The effects of teachers' reference norm orientations on students' implicit theories and academic self-concepts. *Zeitschrift für Pädagogische Psychologie, 31*, 205–219.

Dittrich, E. (2014). Underachievement leading to downgrading at the highest level of secondary education in the netherlands: A longitudinal case study. *Roeper Review, 36*(2), 104–113.

Dronkers, J., & van de Werfhorst, H. (2016). Meritocratisering in schoolloopbanen in nederland. *Meritocratie*, 25–43.

Dupriez, V., & Dumay, X. (2006). Inequalities in school systems: effect of school structure or of society structure? *Comparative Education, 42*(02), 243–260.

Elffers, L. (2018). *De bijlesgeneratie: opkomst van de onderwijscompetitie*. Amsterdam University Press.

Elliot, A. J., & McGregor, H. A. (2001). A 2× 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*(3), 501.

Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion, 32*(3), 158–172.

Felouzis, G., & Charmillot, S. (2013). School tracking and educational inequality: A comparison of 12 education systems in switzerland. *Comparative Education, 49*(2), 181–205.

Fernandez-Cano, A. (2016). A methodological critique of the pisa evaluations. *Relieve, 22*(1), 1–16.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality: Compensation, reinforcement, or neutrality? *American Journal of Sociology, 94*(5), 1146–1183.

Glasersfeld, E.v. (1995). A constructivist approach to teaching. In L. P. Steffe, & J. Gale (Eds.), *Constructivism in education* (pp. 3–15).

Glasersfeld, E.v. (2006). *Radical constructivism in mathematics education* (vol. 7). Springer Science & Business Media.

Gooch, P. W. (2019). *Course Correction: A Map for the Distracted University*. University of Toronto Press.

Gorard, S., & Smith, E. (2004). An international comparison of equity in education systems. *Comparative Education, 40*(1), 15–28.

Groot, A.d., & van Naerssen, R. (1973). *Studietoetsen: construeren, afnemen, analyseren (deel 1)*. Den Haag: De Gruyter Mouton.

Hanselman, P. (2018). Do school learning opportunities compound or compensate for background inequalities? evidence from the case of assignment to effective teachers. *Sociology of Education, 91*(2), 132–158.

Heij, K. (2021). *Van de kat en de bel: Tellen en vertellen met de eindtoets basisonderwijs*. Ph.D. thesis.

Ingenkamp, K. (1977). *Die Fragwürdigkeit der Zensurengebung [The dubiety of school grades]*. Weinheim: Beltz.

Janich, P. (2001). *Logisch-pragmatische Propädeutik: ein Grundkurs im philosophischen Reflektieren*. Velbrück Wiss.

Janich, P. (2015). *Handwerk und Mundwerk: über das Herstellen von Wissen*. CH Beck.

Keeves, J. P., Morgenstern, C., & Saha, L. J. (1991). Educational expansion and equality of opportunity: Evidence from studies conducted by iea in ten countries in 1970–71 and 1983–84. *International Journal of Educational Research, 15*(1), 61–80.

Kolb, A., & Kolb, D. (2018). Eight important things to know about the experiential learning cycle. *Australian Educational Leader, 40*(3), 8–14.

Lago, R., & Beglinger, M. (2009). Schülerjahre. *Wie Kinder besser lernen. München, Zürich (Piper), 4*.

Leeuw, E.d., & Lecina, N. (2016). Onderwijs dat kansrijk maakt. *Technical Report*. Ministerie OCW.

Lubbers, M. J., Van Der Werf, M. P., Kuyper, H., & Hendriks, A. J. (2010). Does homework behavior mediate the relation between personality and academic performance? *Learning and Individual Differences, 20*(3), 203–208.

Marks, G. N., Cresswell, J., & Ainley, J. (2006). Explaining socioeconomic inequalities in student achievement: The role of home and school factors. *Educational Research and Evaluation, 12*(02), 105–128.

McCowan, T. (2016). Three dimensions of equity of access to higher education. *Compare: A Journal of Comparative and International Education, 46*(4), 645–665.

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Cataldi, E. F., Mann, F. B., & Barmer, A. (2019). The condition of education 2019. nces 2019-144. *National Center for Education Statistics*.

McKeachie, W., & Svinicki, M. (2013). *McKeachie's teaching tips*. Cengage Learning.

Merry, M. S. (2019). *Educational Justice: Liberal Ideals, Persistent Inequality, and the Constructive Uses of Critique*. Palgrave McMillan.

Merry, M. S., & Boterman, W. (2020). Educational inequality and state-sponsored elite education: the case of the dutch gymnasium. *Comparative Education*, 1–25.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Chapter 4: Validity in educational assessment. *Review of Research in Education, 30*(1), 109–162.

Muskens, M., & Tholen, R. (2015). Onderzoek sturen op cijfers en rendementen. *Technical Report*. Ministerie van OCW, Research Ned Nijmegen.

Naayer, H., Spithoff, M., Osinga, M., Klitzing, N., Korpershoek, H., & Opdenakker, M. (2016). De overgang van primair naar voortgezet onderwijs in internationaal perspectief. *Technical Report*. GION onderwijs/onderzoek.

Oppers, A.e.a. (2022). De staat van het onderwijs 2022. *Technical Report*. Utrecht: Inspectie van het Onderwijs.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994–1020.

Pillai, S. (2012). Right to education under european convention for the protection of human rights and fundamental freedoms 1950. *Christ University Law Journal, 1*(1), 101–115.

van Pinxteren, M., & de Beer, P. (2016). *Meritocratie: op weg naar een nieuwe klassensamenleving*. Amsterdam University Press.

Plomin, R., & Von Stumm, S. (2018). The new genetics of intelligence. *Nature Reviews Genetics, 19*(3), 148–159.

Popper, K. R. (1971). *The open society and its enemies: The spell of Plato* (vol. 1). Princeton University Press.

Resing, W. C., Elliott, J. G., & Vogelaar, B. (2020). Assessing potential for learning in school children. *Oxford research encyclopedias, education*.

Rheinberg, F. (2001a). Bezugsnormen und schulische Leistungsbeurteilung. *Leistungsmessungen in Schulen, 2*, 59–86.
Rheinberg, F. (2001b). Leistungsbeurteilung im schulalltag: Wozu vergleicht man was womit. *Leistungsmessung in Schulen. Weinheim: Belz*, 59–71.
Richardson, V. (2003). Constructivist pedagogy. *Teachers College Record, 105*(9), 1623–1640.
Rønning, M. (2011). Who benefits from homework assignments? *Economics of Education Review, 30*(1), 55–64.
Ryle, G. (1949). *The concept of mind*. Hutchinson.
Sandel, M. J. (2020). *The tyranny of merit: What's become of the common good?* Penguin UK.
Schultz, K. (2009). *Rethinking classroom participation: Listening to silent voices*. Teachers College Press.
Spiecker, B., & Steutel, J. (2001). Multiculturalism, pillarization and liberal civic education in the netherlands. *International Journal of Educational Research, 35*(3), 293–304.
Terrin, E., & Triventi, M. (2022). The effect of school tracking on student achievement and inequality: A meta-analysis. *Review of Educational Research*.00346543221100850
Tzuriel, D. (2012). Dynamic assessment of learning potential. *Self-directed learning oriented assessments in the asia-pacific* (pp. 235–255). Springer.
United Nations (1948). Universal declaration of human rights. https://www.un.org/en/about-us/universal-declaration-of-human-rights.
Van Rooijen, M., Korpershoek, H., Vugteveen, J., & Opdenakker, M. (2017). De overgang van het basis-naar het voortgezet onderwijs en de verdere schoolloopbaan. *Pedagogische Studiën, 94*(2), 110–134.
Van Vuuren, D., & Van der Wiel, K. (2015). Zittenblijven in het primair en voortgezet onderwijs: Een inventarisatie van de voor- en nadelen. *Technical Report*. Centraal Planbureau.
Vogelzang, M.e.a. (2016). Kansen(on)gelijkheid bij de overgangen PO-VO. Bevindingen en bevorderende en belemmerende factoren. *Technical Report*. Inspectie van het Onderwijs.
Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge university press.
Wolf, R. M. (1998). Validity issues in international assessments. *International Journal of Educational Research, 29*(6), 491–501.
Yang Hansen, K., & Strietholt, R. (2018). Does schooling actually perpetuate educational inequality in mathematics performance? a validity question on the measures of opportunity to learn in pisa. *The International Journal on Mathematics Education, 50*(4), 643–658.
Zapata, J., Pont, B., Figueroa, D. T., Albiser, E., Yee, H. J., Skalde, A., & Fraccola, S. (2014). Education Policy Outlook: Netherlands. *Technical Report*. OECD.
Ziegler, A., & Heller, K. A. (2000). Conceptions of giftedness from a meta-theoretical perspective. *International Handbook of Giftedness and Talent, 2*, 3–21.