The Assessment of Reading Comprehension in English Primary Schools: Investigating the validity of the Key Stage 2 Reading Standard Assessment Test (SAT)

Dr Wayne Tennent, Brunel University, London, UK
College of Business, Arts and Social Sciences, Department of Education

Abstract
The Key Stage 2 Reading Standard Assessment Test (SAT) is a national test of reading comprehension taken by all 11-year olds in state schools in England. The results of this high-stakes test are used for accountability and comparative purposes. While it is acknowledged that these uses have a negative impact on both schools and pupils (Education Commons Select Committee, 2017), less attention has been paid to the validity of the test itself. This article presents a conceptual framework which views assessment as a generic staged process. The framework highlights specific issues related to validity that need to be considered at each stage of the assessment process. When applied to the KS2 Reading SAT this framework shows the test to be of questionable validity and value. Alternative approaches are suggested.

Key words
Reading comprehension, Standard Assessment Tests (SATs), Validity, Key Stage 2, teacher assessment

Introduction
The formal assessment of reading comprehension through tests occurs at two points in primary state schools in England. Known as the Standard Assessment Tests (SATs) these are taken by all children towards the end of Key Stage 1 (KS1) (6-7-year olds) and towards the end of Key Stage 2 (KS2) (10-11-year olds). The tests were introduced in the 1990s to measure attainment (in the core subjects of English, Maths and Science) against a National Curriculum introduced in 1989. This curriculum was divided into age-related key stages, each with specific attainment targets. Initially, the SATs were designed to measure pupil attainment against criterion-referenced levels (Sainsbury, 1996), with the "the purpose of ascertaining what they have achieved in relation to the attainment targets for that stage" (Education Reform Act, 1988, p.2). While schools could take the opportunity to use the results from the SATs to support' school accountability, this was viewed alongside the opportunity to map individual pupil attainment across their school careers (Parliament, 2007).

More recently, the issue of accountability has become the central focus. The rationale for the SATs is more obviously driven by Government policy undertaken for the purposes of individual school accountability (Wiliam, 2010), and comparison between schools (Newton, 2009). This is most noticeable with the KS2 SATs where results are published and schools are placed in 'league table' order.

The effects of this measurement culture are felt at an individual school level. In their report on *Primary Assessment*, the parliamentary Education Commons Select Committee (2017) investigated these effects. In a section entitled 'Impact of high-stakes assessment', it notes

'the pressure that schools are under to achieve results at Key Stage 2' (p.16) and the negative effect this can have on children's well-being. A further consequence of high-stakes testing is for the curriculum to become separated from the assessment process (Moss, 2017). This impacts upon classroom practice as the curriculum becomes notional; and attention switches to focus on those subjects being tested. Wiliam (2010) notes that this narrowing process affects how practitioners view the construct being tested. For example, the construct being tested in the KS2 Reading SAT paper is reading comprehension. The suggestion is that teachers view reading comprehension solely through the prism of the test demands, and are likely to 'teach children whatever the test asks of them' (Moss, 2017, p.62). Effectively, teachers 'teach to the test' – a phenomenon also acknowledged in the same Education Select Committee report.

There is a parliamentary precedent then, to describe the KS2 SATs as high-stakes tests, which have potentially negative personal and emotional consequences for schools, teachers and children; which lead to a narrow, pressurised curriculum; and where teaching to the test is common.

The Select Committee however, was more accepting of the tests themselves. The Standards and Testing Agency (STA), the body responsible for developing the SATs, reported to the Select Committee that they were satisfied with the tests. Responding to widespread criticism that the 2016 KS2 Reading SAT was too difficult, the STA's Chief Executive stated that the test 'did perform as… expected' (p.9) as it had a 'sufficient' (p.10) spread of scores at the lower and the higher ends. By suggesting that the test performed as 'expected' a claim to validity is implied; and by noting a 'sufficient' spread of scores a nod to a normal distribution of scores across a cohort is also implied. By implication then the SATs are linked to the field of psychological measurement. The link becomes explicit when the language around the test is considered. As noted above, Wiliam (2010) talks about the testing of *constructs*. Baird et al. (2017) point to Cronbach and Meehl (1955) who defined a psychological construct as 'some postulated attribute of people, assumed to be reflected in test performance' (p. 283). Indeed, Baird et al. explicitly link the terms 'construct', 'testing' and 'psychology'.
This has been reaffirmed with the advent of an updated National Curriculum (2014) where assessment against criterion-based levels has been removed. Children taking the SATs are now judged against a process of scaled scoring. Providing each child with a scaled score (discussed in more detail later) allows for children to be judged against each other and to be placed in a type of rank order. As such, scores in these tests are being viewed normatively. The reasons for this are likely to political. As Tengburg (2017) notes the desire to make comparisons on both national and international levels has led to a move towards standardised measurements of learning, of the type used in psychological testing.
Essentially, the KS2 Reading SAT aims to measure the construct of reading comprehension through the use of a psychological test. This is problematic because the SATs takes place in an educational context, and as Baird et al (2017) note, assessment in Psychology and Education are not necessarily viewed in the same way. Assessment in Education is invariably linked with learning to inform future teaching. This is most clearly seen in the process of Assessment for Learning (AfL) (Black & Wiliam, 1998) which promotes the use of assessments for formative purposes. Psychological based assessments in an educational context involves creating measurement tools in the form of tests with the results used to discover or support normative trends, as used in the SATs.

This raises the issue of what constitutes acceptable evidence of learning. In the context of reading comprehension, for example, Ellis and Smith (2017) have proposed a multi-faceted conceptual framework from which to consider reading comprehension and its assessment. It acknowledges the cognitive and meta-cognitive aspects of reading, but also acknowledges the reader's social and cultural capital, and their identity. Ellis and Smith suggest that to establish a rounded understanding of a child's reading comprehension abilities, evidence from these latter dimensions must be considered. Clearly, these latter dimensions are difficult – if not impossible – to standardise, and as a consequence are not considered in the current approach to assessment. The danger here is that the construct being measured becomes hollowed out to only those aspects that lend themselves to measurement. Indeed, Baird et al (2017) note that in many instances the construct being measured in an educational context is not sharply conceptualised.

Performance on the KS2 reading SAT provides the *only* evidence of children's reading comprehension ability at the end of primary education in England. This evidence is now primarily used for school accountability purposes so one would expect the test to perform this function rigorously. Given that the SATs aim to assess a supposedly measurable construct in a test situation, and assertions have been made by the STA which state that the KS2 Reading SAT performs to expectations, investigating the validity of the test seems a reasonable starting point. This paper will consider the validity of the KS2 Reading paper, its construction, and how its results are used.

Validity: Issues with definition and explication

As with assessment, the concept of validity will not necessarily be viewed in the same way in the fields of Psychology and Education. The root cause of this difference is one of theoretical perspective (Crotty, 1998). The field of Education includes interpretivist and critical perspectives. Sociological and political positions are likely to be considered relevant alongside (or for some, instead of) psychological ones. Referring again to the Ellis and Smith (2017) framework for reading, it would be considered appropriate to explore social and cultural phenomena despite these being situation specific. The field of Psychology in general is more closely aligned to positivist perspectives and the 'scientific' approach with the aim of adopting a neutral, objective stance. . As the KS2 reading SAT reflects this psychological perspective, the scope of this discussion on validity will focus on the area of psychological assessment in education alone.

For those concerned with educational and psychological measurement, validity has been described as the 'most fundamental consideration in developing tests and evaluating tests' (AERA, APA, & NCME, 2014). It is 'the most important criterion' (Koretz, 2008, p.215) in test evaluation. Indeed, Newton and Baird (2016) state that validity is 'the most important term in the educational and psychological measurement lexicon' (p.173). Given this importance, it is perhaps surprising that in over 100 years no consensus definition of the term has been reached (Newton & Shaw, 2016). Newton and Shaw (2014) located 151 different variants of validity in the literature. While there is significant overlap between some of these variants, the fact that they have been assigned different labels at all suggests why there are likely to be variations in definition. Indeed, Cizek (2016) goes as far as to say that many authors actively avoid any attempt to define validity. The issue is not one simply of definition however, but rather how the term is conceptualised. Kane (2016), for example,

states that rather than focus on defining validity it is more important to explicate the term. Explication involves making the concept of validity clearer, thus increasing its usefulness.

Explication allows for consensus points to be located. Cizek (2016,) locates a number of consensus points in the debate on validity. These include – but are not restricted to – the following points which state that validity pertains to:

- 'The intended inferences or interpretations made from test scores and not to the tests themselves' (p.213);
- How these interpretations relate to the specific construct being tested;

and that,

- The 'Beliefs, assumptions and values' (p.213) of evaluators mean that the same results can be interpreted in different ways;
- There needs to be a unitary view of validity, rather than the multiple types of 'validities' located by Newton and Shaw.

Explication, however, also allows for points of contention to be noted. The most obvious point of contention relates to the need for validity to be viewed as a unitary construct. The problem here is that there is no consensus on the scope of the term. The difficulty centres on whether there is an ethical (Messick, 1989) element to validity which requires the *use (or uses)* of assessment to be considered. For Kane (2016), the matter of test score interpretation – and the meanings drawn from them – should not be considered in isolation from the how the scores are then used. Newton and Shaw (2014) suggest the two are intrinsically linked. The purpose of attempting to accurately measure a construct (such as reading comprehension, for example) through a test, is so that something can be done with the results. For Cizek (2016), the concept of validity cannot be stretched this far: the interpretation of test scores and the use to which they are eventually put, are two separate issues that should not be conflated. Instead, Cizek links the uses of test results to a process of justification. Aligning validity with 'psychometric traditions' (p.218) related to the interpretation of score meanings, he argues that the uses of test scores do indeed need to be justified, but that this is beyond the remit of test validation.

Two definitions presented by Newton and Shaw (2016) capture these differing perspectives:

- A test is said to be valid if it measures what it purports to measure. (Kline, 1998, p. 34)
- Validity is the adequacy and appropriateness of the interpretations and uses of assessment results. (Miller, Linn, & Gronlund, 2009, p. 70)

The first definition takes a narrower view of validity. The latter takes the wider view which encompasses the assessment use.

As noted in the introduction, any position taken on validity is likely to relate to whether a psychological or an educational perspective is assumed. To exemplify the point Newton and Baird (2016) note how psychometric scientists and educational practitioners would view validity differently. Psychometric scientists are likely to take a narrower conception of validity, similar to Cizek's (2016) view, while education practitioners are likely to favour a conception that encompasses ethics, in line with Kane's (2016) view. In a similar vein, Newton and Shaw (2016) characterise these differing stances as *conservative* and *liberal*

(*conservative* relating to the narrow view and *liberal* to the broader view). While the terms *conservative* and *liberal* are not capitalised, they are certainly loaded, and it is not difficult to see how epistemological assumptions can turn into ideological positions. According to these theorists, educators would more obviously position themselves as liberals, working from the assumption that the uses of assessment should be a facet of validity.

Validity in the assessment process: Developing a theoretical framework

The validity definition debate might appear to an academic argument over semantics and concepts. However, the high-stakes KS2 Reading SAT is rooted in psychological measurement processes and validity is central to this. It is necessary then to consider both the narrow and wider perspectives of validity to make a rounded comment on the efficacy of this test, as well as to consider where educationalists should position themselves in the validity debate. A theoretical framework which encompasses both views is required. Newton (2007) suggests that the purpose of any assessment act in education can be viewed on three levels: The judgement level, the decision level and the impact level. By combining Cizek's (2016) consensus points and Newton's levels of assessment, it is possible to create a theoretical framework from which to analyse the validity of the KS2 reading SAT. The framework views assessment as a process (See Figure 1) and raises key questions at each stage that must be considered.

[Figure 1 near here]

Stage 1: The Assessment Aim

The beginning of the assessment process links to what Newton (2007) describes as the *judgement level*. The judgement level relates to the *aim* of assessment act. All assessment acts have an aim whether it relates to a formal test or an ongoing classroom assessment. However, this aim does not exist in a vacuum. One of Cizek's consensus points is that the interpretation of scores should relate to the construct being assessed. This implies that the construct itself needs to be clearly theorised and explicit.

Stage 2: The Assessment Event

The assessment event describes the point in the assessment process where evidence is collected and interpreted. Cizek notes that beliefs and values can play a part in the interpretation of results. The collection of evidence also requires some sort of assessment tool, such as a test. While there is consensus that validity is concerned with the interpretation of the test and *not the test itself* (Newton, 2012; Cizek, 2016), this assumption is questioned later in the article, specifically in relation to the testing of reading comprehension.

Stage 3: The Assessment Use

This stage relates to the Newton's decision *level.* The decision level relates to how the assessment judgement is *used*. It is at this point that the wider concept of validity becomes relevant, and the question arises of whether the ethical use of tests should be a facet of the term.

Stage 4: The Assessment Impact

This final stage relates to Newton's *impact level of assessment*. Harlen (2007) agrees with Newton that assessment events have an impact; and indeed, she describes impact as one of the central principles of assessment. Impact relates to the consequences of using an assessment in a specific way. The use of assessment results is neither passive nor neutral.

Impact is not a facet of validity per se. As shown later in this article, the impact of an assessment use can be both expected and unexpected. As such, the impact of an assessment use cannot always be anticipated. Impact is included in this framework because it provides an end (or reflection) point to the wider assessment process.

The KS2 Reading SAT as an assessment process

This assessment process can be applied to the KS2 Reading SAT. The stated *aim* of the test is to 'ascertain what pupils have achieved in relation to the attainment targets outlined in the national curriculum' (STA 2018a, p.6) for reading. The *assessment event* takes the form of a one-hour test featuring a booklet of unseen texts (of increasing complexity) and an answer booklet. The test is written although very few questions require extended responses. The mark scheme for the 2019 Reading SAT states that the test assesses 'aspects of comprehension that lend themselves to a paper test' (STA, p.3). This is a statement that appears year-on-year. In terms of the *assessment use*, it is accepted practice for data from the Reading SAT to be put to multiple uses. The results are to be used to hold schools to account for attainment and progress in reading; to inform parents and other schools about individual performance; and to act as a benchmark between schools both locally and nationally (STA, 2018a). These uses are about reporting, accountability, and the summation of learning.

Newton (2007) argues that the uses of national test results go far beyond those stated –and they are not necessarily ones which might been anticipated by test designers and policy makers. He locates 22 different uses of assessment results generally. Some of these do map onto the stated uses of the SATs. As part of the institution monitoring process for example, schools are ranked in 'league tables' for comparative purposes; and in terms of accountability, if attainment is below 'expected' levels a school inspection might be triggered. Secondary schools might use the results for placement purposes such as streaming children into appropriate classes on entry – but they are not compelled to do so. Informing parents about individual performance is an action taken by schools; what parents actually do with this information is somewhat opaque. Newton notes that other parents might use the results of the SATs to inform school choice decisions for their own children. Whether this is an intended use is debatable. The potential *impact* of these uses can be both positive and negative, some of which could be evidenced and some of which are likely to involve level of speculation.

Figure 2 maps the 'official' assessment process for SATs encompassing the narrow and wide views of validity. The point to make here is that further *unintended* impacts are possible.

[Figure 2 near here]

Investigating the validity of the Reading SATs from the narrow view: Fundamental issues

There are some fundamental issues which challenge the validity of the KS2 SAT Reading test when considered against the narrow view of validity. These relate to the key questions outlined in Figure 1 for the assessment aim and the assessment event. The most obvious issue relates to the statement that the test measures 'aspects of comprehension that lend themselves to a paper test'. This acknowledges that specific other aspects of comprehension are missing. This immediately restricts any claim that the test provides evidence of achievement across the entire reading curriculum, which is the stated aim. Clearly, this limits the effectiveness of the assessment tool (The test) and compromises any interpretation of the test score. There are a number of further issues.

1. The Construct of Reading Comprehension is weakly theorised

The curriculum which informs the teaching of reading in England is underpinned by the simple view of reading (Gough & Tunmer, 1986). This is a psychologically based framework which has provided the conceptual basis for the teaching of reading in England since 2006 (Rose, 2006). The simple view of reading states that there are two overarching components to reading: word recognition (Decoding) and spoken language comprehension. This is often presented as a hypothetical formula:

Reading = Word recognition x Spoken language comprehension

Describing reading as a multiplicative process indicates that both word recognition and comprehension are required. Research linked to the simple view has shown that the two processes can be disassociated (Catts et al. 2005), which in turn allows for specific reading difficulties to be located (Aaron et al., 1999). Beyond this, the construct of comprehension is weakly theorised.

Spoken language comprehension is not the same as reading comprehension

The simple view of reading refers to spoken language comprehension only. This is not the same as reading comprehension. Cunningham (2005) notes that oral and written language are very different methods of communication. Readers are required to engage with different syntactical and text structures which are not found in ordinarily in speech (Concannon-Gibney & Murphy, 2010). Indeed, readers develop their spoken language through engagement with texts (Perfetti et. 2005). The focus on spoken language comprehension may have some credence in relation to early readers but those pupils taking the KS2 Reading SAT are, on the whole, likely to be relatively proficient. They will be acquiring language from written texts. The construct presented in the simple view of reading does not acknowledge nor account for this.

Comprehension is componential

Duke (2005) has categorically stated that comprehension is 'not a unitary construct', but rather that text comprehension is componential. It is composed of numerous processes and factors that work interactively. Tennent (2015) has presented a framework to demonstrate this; these include linguistic processes, cognitive processes, metacognitive processes and knowledge factors. Indeed, Kintsch and Kintsch (2005) state that it is not possible to measure comprehension as a uniform construct. Text comprehension is a complex process and tests of 'comprehension' cannot easily separate these components. Kendeou et al. (2016), for example, note that high-stakes tests developed for use in the United States do not adequately assess the centrally important component of inference making. Even a relatively straightforward question such as, 'What does the word *x* mean?' can be answered through vocabulary knowledge without reference to the text; inference making at the sentence level through the use of context; or even morphological knowledge. The simple view of reading does not acknowledge the componential nature of comprehension, nor its complexity.

A list of reading skills is not a construct

The KS2 Reading SAT mark scheme (2019) provides an indication of what the test attempts to cover. This links to the current, updated National Curriculum (2014) for Reading. This describes comprehension as a domain which requires the reader to do the following:

1. give / explain the meaning of words in context;
2. retrieve and record information / identify key details from fiction and non-fiction;
3. summarise main ideas from more than one paragraph;
4. make inferences from the text / explain and justify inferences with evidence from the text;
5. predict what might happen from details stated and implied;
6. explain how information/ narrative content is related and contributes to meaning;
7. explain how meaning is enhanced through choice of words and phrases;
8. make comparisons within the text.

This is not a construct but rather a set of desired reading skills; and these reading skills are not necessarily straightforward to isolate. As noted above Skill 1 (explaining word meanings) might involve inference making, or the word might just be known. Skill 3 (summarising) might depend upon the inferences made. And indeed, Skill 5 (predicting) is a type of inference (Tennent, 2015). Skill 7 (how meaning is enhanced through word choice) refers to the literariness (Zwaan, 1996) of a text. Some might argue that literary criticism is not a facet of comprehension, so how comprehension is theorised becomes critically important.

In the 2019 Reading SAT 29 of the 50 marks were described as addressing Skill 2 (retrieving information) and Skill 4 (making inferences). As the mark scheme states the test was developed to assess 'the aspects of comprehension that lend themselves to a paper test'. The SAT cannot be said to assessing comprehension as a whole, even if this list of reading skills was accepted as a construct. At no point are links made back to the supposed conceptual framework of the simple view of reading.

No developmental path for comprehension has been mapped

The scores achieved by children in the KS2 Reading SAT are interpreted to describe them as being *below*, *at*, or *above* 'expected' levels. These categorisations are not criterion-referenced, but rather linked to scaled scores. When scores are scaled, the raw score – the

score actually achieved in the KS2 Reading test out of 50 marks, for example – is statistically adjusted through comparison with test scores achieved by previous cohorts. As such, scaled scores attempt to address invariance (or generalisability) between tests. The Department for Education (2019) actually state that the use of scaled scores allows differences in test difficulty between years to be accounted for. In 2019, a raw score of 27 would give a scaled score of 99, which is considered to be *below* the expected level. A score of 28 would give a scaled score of 100, which is considered to be *at* the expected level.

The categorisations of *below*, *at* and *above* the expected levels are thus an attempt at norm-referencing. This is problematic because no developmental pathway has been mapped for reading comprehension. There is no evidence to show how comprehension looks different between a 7-year old and an 8-year old, for example. It is for this reason that Kintsch and Kintsch (2005) state that there is little value in putting a score to comprehension. As such, scores in a reading comprehension test are likely to tell us very little about an individual child's reading comprehension. Indeed, the difficulty here is that there is no research informed description of what 'expected levels' look like at the age of 11. No score – raw or scaled – can capture the complexity of text comprehension because the components noted above work both uniquely and interactively. Given this, it is difficult to see the value of generating normative data because how these components are accessed by readers is going to vary. This makes the stated aim of the Key Stage 1 & 2 tests – to ascertain what pupils have achieved – impossible to achieve through normative means.

2. The Test itself does matter

As noted above Newton (2012) and Cizek (2016) both state that when we consider validity in assessment, it is not about the actual test itself, but rather how the results are interpreted. This is a consensus point but is questionable when testing reading comprehension. There is evidence to suggest that different tests of reading comprehension "tend to measure different aspects of reading comprehension" (Tengburg, 2017, p.85). According to Tengburg, there are two reasons for this. The first acknowledges the point that different tests are likely to measure different skills, and indeed, the KS2 SAT is an example of this. If there is clarity as to what is being assessed, then any inferences made from scores should relate to those skills being assessed only. In this instance then, the focus can be on the inferences drawn, and not the test. The second reason for differences in tests of reading relate to what Tengburg describes as the 'contextual factors' and includes text type and the topic of the text. These contextual factors are critically important because they are likely to impact upon a reader's response. In tests which aim for a measure of standardisation through scores this is never acknowledged in the inferences drawn.

Test developers work within a social and cultural context

Baird et al. (2017) state that, 'test features may reflect and amplify aspects of the culture in which they arise' (p.322). This should not be a surprise given that all tests, regardless of the construct being assessed, are designed by people. As Harkins and Singer (2009) state these people are likely to write in relation to their own cultural background. Indeed, these authors cite Christensen (2000) whose research showed that large-scale tests 'typically measure middle- or upper-class experiences' (p.80). Florez Petour (2017) notes that test designers have to make a political decision when deciding about the underlying theories that inform a

test. They have a choice as to whether they follow the views of policy makers. Whether they choose to or not is secondary in this discussion to the fact that they do have this choice. This links to another of the consensus points. If there is consensus that the 'beliefs, assumptions and values' can impact upon test interpretation, the 'beliefs, assumptions and values' of those designing the test must also be a factor. It is illogical to suggest otherwise.

Comprehension tests are mediated by text(s) which cannot control for knowledge
Knowledge bases are a factor which inhibit or facilitate the ability of the children to demonstrate their understanding of any text. The question is whether this is accounted (or controlled) for in testing situations. A case for this can be made in some instances. If, for example, a pupil is completing a History test, or a Science test it can be assumed that the curriculum has been covered and, to some extent, there is at least the possibility of stating that pupils have a similar knowledge.

This does not work for reading comprehension tests currently. Every test of reading comprehension requires the reader to engage with written text; and as such, their understanding of each text will be mediated by what they can bring to it from their knowledge and experiences. The KS2 Reading SAT requires readers to engage with *unseen* texts on *unidentified* content. This is problematic. Arguing for a non-unitary definition approach to comprehension testing, Wixson (2017) notes that performance on reading comprehension tests will be affected by background knowledge and passage familiarity. As the texts used in the Reading SATs are unseen with unidentified content, some children are likely to have an advantage (and conversely some will be disadvantaged) based on their background knowledge and experiences. The controversial 2016 KS2 Reading SAT for example featured a passage on Dodos. If as a 10-11-year-old reader you have an interest in extinct birds your ability to engage with the text is likely to be facilitated – and the ability to answer questions might simply be linked to domain knowledge and personal experiences.

It is for this reason that invariance (or generalisability), once again, will be an issue. Scores taken from one year cannot necessarily be treated the same way the next (Kane, 2016). The DFE (2019) specifically link the use of scaled scores to the fact that the questions which children are required to answer are different year-on-year. They make no mention of the texts used. The reason why the questions are different in the KS2 Reading SAT is because the texts used are not the same. Indeed, Tennent et al. (2008) noted the lack of year-on-year consistency with this test, and one of the reasons for this was the variety of texts used. All texts are culturally and socially located, which explains why readers approach each text with unique knowledge bases. The assumption that the texts used are in some way neutral cannot be substantiated.

The text matters – and therefore so does the text used in a test.

The KS2 Reading SATs in the wider view of validity
The wider view of validity considers the use and impact of assessments. In relation to the KS2 SATs some further unintended uses and impacts can be located beyond those stated in the documentation, and reflect some of the 22 assessment uses noted by Newton (2007). These are shown in Figure 3.

[Figure 3 near here]

Again, these uses and impacts can be both positive or negative, and again some of the impacts are speculative. For example, Year 6 teachers might analyse the results of one year's test results to inform and adjust their teaching for the next cohort. The impact of this formative assessment should ensure more effective teaching. This could be considered as a positive use and impact. A less positive use and impact relates to Valuation where there is an impact on house prices. This is not speculative as there is evidence to show that house prices have increased in catchment areas where schools perform well on KS2 SAT tests (Battistin & Neri, 2017). The impact of this is that schools are likely to become economically and socially homogenous.

Yet in all the circumstances outlined here, the uses (and potential associated impacts) – intended or otherwise – occur in relation to data gathered by a test which does not adequately measure the construct it claims to. So, teachers using the results of tests formatively to inform future teaching becomes a less useful practice, given the year-on-year issues with invariance because of the different texts used. House prices may increase where SATs results are better, but this does not necessarily mean 'better' readers. It might just mean more intensive 'teaching-to-the-test' in a narrow curriculum where the construct of reading comprehension is viewed through test demands only. And secondary schools might be over- or under-estimating reading ability if using the results of the Reading SAT for placement purposes, when the scores achieved are of limited use.

Addressing some of the difficulties with the reading SAT

There is then a lack of clarity and alignment across the assessment process with regard to the KS2 Reading SAT which attests to significant issues with the validity of the test. These stem from the inability to address the consensus points which underpin the narrow view of validity. If the Reading SAT is to be retained, it is necessary to take different approaches to alleviate some of the central issues.

Taking a topic-based approach to the texts used

The Education Select Committee have acknowledged that the high-stakes culture around SATs causes the curriculum to be narrowed and teaching-to-the-test becomes normalised. The analysis in this paper notes that for the KS2 Reading SAT numerous further problems exist. One of these centres around the texts used. They are unseen with unidentified content. As such, they do not account for – or indeed even consider – variations in background knowledge. These issues can be addressed (to some extent) by a change in approach whereby a topic (with some guidance as to possible content) is presented to schools at the start for the academic year, such as 'Extinct Animals'. This topic can then form the basis of the texts to be used in the Reading SAT later in the year. Teachers could then develop a curriculum in relation to this topic. They might be teaching 'for' the test, but not necessarily 'to' the test. They would also be using pedagogical approaches which are not fundamentally aimed at test preparation. If the topic is interesting, children are likely to be more engaged and feel less pressurised. This would be beneficial to their well-being. Outlining a topic would not completely control for background knowledge but it would go some way towards addressing

the issue. The texts could still be unseen if commissioned specifically for the test and it would allow well-known contemporary authors to be involved. This would ensure quality and relevance. Given that the STA's net expenditure on assessment and policy development for 2016-17 was £54,120,000 (STA, 2018b), this would be a relatively small expense.

Teacher assessment

The KS2 Reading SAT attempts to measure only those 'aspects of comprehension that lend themselves to a paper test' (STA, 2019, p.3). This statement in itself acknowledges that reading comprehension is inadequately assessed, and that evidence is required from other sources. The most obvious place is the classroom through teacher assessment of everyday learning. The Ellis and Smith (2017) framework outlined earlier provides an obvious starting point. As noted, this framework acknowledges the cognitive aspects of reading, but also acknowledges the reader's social and cultural capital, and their identity. This would encompass the knowledge and experiences readers bring to the reading act, allowing teachers to consider the role of the text in the assessment, and indeed the variety of texts which their pupils are engaging with. This rounded view allows for a deeper inferential analysis of a child's reading profile while also moving away from meaningless inferences such as 'expected level'.

Such approaches rely on practitioners' assessment literacy (Stiggins, 1991). As a central component of teaching and learning, practitioners need to be 'literate' in the language and discourse of assessment. For example, one could argue that the concept of validity is not really discussed by practitioners; indeed, the term seems to be the preserve of academic discourse. This should not be the case. As Pastore and Andrade (2019) note, assessment literacy does not simply apply to understanding standardised testing. The key questions described in the assessment process here shows that it applies to everyday classroom assessment. Practitioners, for example, should be empowered to consider whether the construct they are testing is clearly conceptualised.

To re-iterate, the KS2 SAT provides the only evidence of reading ability at 11 years old; and as the associated documentation acknowledges this evidence is partial because it only covers those parts of the curriculum which lend themselves to a written test. Teacher assessment would thus seem essential.

Taking this further, developing assessment literacy would negate the need for national testing in the long run.

Conclusion

The assessment process presented in this paper provides a way of explicating the concept of validity. Asking key questions at each stage of the process allows for both the narrow and wider views of validity to be examined. In relation to the KS2 Reading SAT, the construct of comprehension is poorly conceptualised (supporting Baird et al.'s (2017) point outlined in the introduction); the test design fails to control for background knowledge; the inferences made from the test cannot be substantiated; and as a consequence, all uses – both intended and unintended – are taken on false premises.

Whether the assessment use should be considered a facet of validity (a second point by Baird et al. raised in the introduction) is a longer-term discussion; and depends upon how educationalists engage with the epistemological assumptions around questions of test validity. As noted above, those in the field of education have been positioned by those in the psychological and educational assessment community as being liberal advocates of the wider view of validity. It would not be surprising if educators did indeed associate tests such as the KS2 reading SAT with how their results are used. The results of national tests are used in numerous ways, often for overtly political purposes. For example, the year-on-year improvement in the results of the phonics screening check – a high-stakes national test of phonic knowledge taken by all 6 years olds in state schools in England – led one Government minister (Gibb, 2015) to suggest that this improvement in scores demonstrates an improvement in general reading. To be clear, there is no evidence to support this assertion. It is not a test of general reading, and indeed, an NFER (2015) evaluation, commissioned by the DFE and published three months before the minister's statement, found that there was no evidence to suggest that improvement in phonic decoding had led to any general improvement in wider reading skills.

When claims such as these are made, responding to them does involve taking an ethical stance. Whether this should be perceived as an issue of validity is debatable, but may actually be missing the point. The issue is more about ensuring that the results of assessments are used responsibly. Indeed, Florez Petour's (2017) argument that test designers have agency in how they develop tests implies a request for them to take responsibility for how the tests they develop are used. A further implication of Florez Petour's argument is that test designers seem reluctant to do so. Colleagues from the psychological and educational testing community do not seem to have responded to the above claim made by the government minister, for example. Perhaps this goes some way in capturing the difference between the so-called conservative and liberal positions.

So, educationalists who engage with reading comprehension assessment, and who are concerned about the responsible use of SAT scores do need to take make their position clear on this. A more fundamental issue is that the concept of validity is tied to psychological measurement, and the associated measurement tools reflect psychology's epistemological assumptions. To repeat, there is little to be gained in putting a score to comprehension. The analysis here would suggest that the measures do not work, and the assumptions on which they are based are inappropriate. Adjusting the test and promoting teacher assessment, based on the model promoted by Ellis and Smith (2017) for example, would provide some remediating actions. In the wider context of the national testing of reading comprehension though, those involved and interested in developing children's reading comprehension might want to separate themselves from the ineffective practices, i.e., those based in psychology, which are currently used.

Word Count: 6642

Disclosure statement

There are no conflicts of interest

ORCID

Wayne Tennent https://orcid.org/0000-0001-7868-4877

References

Aaron, P.G., Joshi, R.M. and Williams, K. 1999. "Not all reading disabilities are alike." *Journal of Learning Disabilities* 32: 120–37.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, NCME]. 2014. *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Baird, J.O., Andrich, D., Hopfenbeck, T. N., and Stobart, G. 2017. "Assessment and learning: fields apart?" *Assessment in Education: Principles, Policy & Practice* 24 (3): 317-350.

Battistin, E. & Neri, L. 2017. "School Performance, Score Inflation and Economic Geography." *IZA Discussion Papers 11161, Institute of Labor Economics (IZA).*

Black, P. & William, D. 1998. Inside the Black Box. London: Kings College.

Brookhart, S. M. 2011 . "Educational assessment knowledge and skills for teachers." *Educational Measurement: Issues and Practice,* 30: 3 - 12.

Catts, H.W., Hogan, T.P. and Adlof, S.M. 2005. "Developing changes in reading and reading disabilities." In *The Connections between Language and Reading Disabilities*, by H.W. Catts and A.G. Kahmi (eds), 50–71. Mahwah, NJ: Erlbaum.

Christensen, L. 2000. *Reading, writing, and rising up: Teaching about social justice and the power of the written word.* Milwaukee, WI: Rethinking Schools.

Cizek, G.J. 2016. "Validating test score meaning and defending test score use: different aims, different method." *Assessment in Education: Principles, Policy & Practice,* 23 (2): 212-225.

Committee, Education Commons Select. 2017. *Primary Assessment.* Accessed 7 29, 19. https://publications.parliament.uk/pa/cm201617/cmselect/cmeduc/682/682.pdf .

Concannon-Gibney, T. and Murphy, B. 2010. "Reading practice in Irish primary classrooms: too simple a view of reading? ." *Literacy,* 44 (3): 122-130.

Cronbach, L., and Meehl, P. E. 1955. "Construct validity in psychological tests." *Psychological Bulletin, 52,* 281–302.

Cunningham, A.E. 2005. "Vocabulary growth through independent reading and reading aloud to children." In *Teaching and Learning Vocabulary: Bringing Research to Practice*, by E.H. Hiebert and M.L. Kamil (eds), 45–68. Mahwah, NJ: Erlbaum.

Duke, N.K. 2005 . "Comprehension of what for what: comprehension as a non-unitary construct." In *Current Issues in Reading Comprehension and Assessment*, by S. Paris and S. Stahl (eds), 93–104. Mahwah, NJ: Erlbaum.

Education, Department For. 2014. *National Curriculum in England: Framework for Key Stages 1 to 4.* Accessed 10 10, 2019. www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4 .

Education, Department for. 2018 . *National curriculum test handbook: 2018 Key stages 1 and 2* . Accessed 10 10, 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765749/2018_NCT_Handbook_PDFA.pdf.

Education, Department For. 2019. *Understanding scaled scores at key stage 2.* Accessed 7 29, 2019. https://www.gov.uk/guidance/understanding-scaled-scores-at-key-stage-2.

Education Reform Act. 1988. https://www.legislation.gov.uk/ukpga/1988/40/pdfs/ukpga_19880040_en.pdf

Ellis, S., and Smith V. 2017. "Assessment, teacher education and the emergence of professional expertise." *Literacy,* 51 (2): 84-93.

Flórez Petour, M. T. 2017. "Power, ideology, politics: 'the elephant in the room' in the relationship between assessment and learning." *Assessment in Education: Principles, Policy & Practic*e, 24 (3): 433-439.

G., Moss. 2017 . "Assessment, accountability and the literacy curriculum: reimagining the future in the light of the past." *Literacy* 51 (2): 56-63.

Gibb, N. 2015. *Focus on phonics vindicated by results. .* Accessed 9 15, 2019. https://www.nickgibb.org.uk/news/focus-phonics-vindicated-results.

Gough, P.B. and Tunmer, W.E. 1986. "Decoding, reading and reading ability." *Remedial and Special Education, 7* 6–10.

Harkins M. J. and Singer, S. 2009. "The Conundrum of Large Scale Standardized Testing: Making Sure Every Student Counts." *Journal of thought,* Spring-Summer: 77 -90.

Harlen, W. 2007. *Assessment of Learning.* London: Sage.

Kane, M. T. 2016 . "Explicating validity." *Assessment in Education: Principles, Policy & Practice,* 23 (2): 198-21.

Kendeou, P., McMaster, K. L. and Christ, T. J. 2016. "Reading Comprehension: Core Components and Processes." *Policy Insights from the Behavioral and Brain Sciences,* 3:1 62 –69.

Kintsch, W. and Kintsch, E. 2005. "Comprehension." In *Current Issues in Reading Comprehension and Assessment.* , by S.G. Paris and S.A. Stahl (eds), 71–92. Mahwah, NJ: Lawrence Erlbaum Associates.

Kline, P. 1998 . *The new psychometrics: Science, psychology and measurement.* London: Routledge.

Koretz, D. 2008 . *Measuring Up: What Educational testing really Tells Us.* Cambridge, Mass.: Harvard University Press.

Messick, S. 1989. "Validity." In *Educational measurement (3rd ed.)*, by R. L. Linn (ed.), 13–103. New York, NY: American Council on Education & Macmillan.

Miller, M. D., Linn, R. L., and Gronlund, N. E. 2009 . *Measurement and assessment in teaching (10th ed.).* Upper Saddle River, NJ: Pearson Education.

Newton P.E. and Baird, J. 2016. "The great validity debate." *Assessment in Education: Principles, Policy & Practice,* 23 (2): 173-17.

Newton, P.E. and Shaw, S.D. 2014. *Validity in Educational and Psychological Assessment.* London: Sage.

Newton, P.E. 2007. "Clarifying the purposes of educational assessment." *Assessment in Education,* 14 (2): 149–170 .

Newton, P.E. 2009. "The reliability of results from national curriculum testing in England." *Educational Research,* 51 (2): 181-212.

NFER. 2015. *Phonics screening Check Evaluation Final report Research report.* June . Accessed 10 10, 2019. https://www.nfer.ac.uk/publications/YOPC03/YOPC03.pdf.

Newton, P.E. 2007. *Evaluating Assessment Systems. QCA.* Accessed 8 1, 19. https://dera.ioe.ac.uk/18993/1/Evaluating_Assessment_Systems1.pdf.

Newton, P.E. 2012. "Clarifying the Consensus Definition of Validity." *Measurement, 10* 1–29.

Parliament. 2007. The evolution of the National Curriculum: from Butler to Balls. https://publications.parliament.uk/pa/cm200809/cmselect/cmchilsch/344/34405.htm

Pastore, S. and Andrade, H. 2019 . "Teacher assessment literacy: A three-dimensional model." *Teaching and Teacher Education,* 84: 128 - 138.

Perfetti, C.A., Marron, M.A. and Foltz, P.W. 1996. "Sources of comprehension failure: theoretical perspectives and case studies." In *Reading Comprehension Difficulties: Processes and Intervention.*, by C. Cornoldi and J. Oakhill (eds), 137-165. Mahwah, NJ: Lawrence Erlbaum Associates.

Rose, J. 2006. *Independent Review of the Teaching of Early Reading: Final Report.* Accessed 10 10, 19. https://dera.ioe.ac.uk/5551/2/report.pdf.

Sainsbury, M. 1996 "Assessing English". In *SATs: The Inside Story: The Development of the Frst National Tests for Seven-year-olds by M. Sainsbury (ed), 16-32. Slough: NFER.*

STA. 2018a. *National curriculum Test Handbook: 2018 Key stages 1 and 2.* Accessed 10 10, 19. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765749/2018_NCT_Handbook_PDFA.pdf.

STA. 2019 . *National Curriculum Tests - English Reading Test Mark Schemes.* Accessed 10 10, 19. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/803889/STA198212e_2019_ks2_English_reading_Mark_schemes.pdf.

STA. 2018b. *Standards and Testing Agency Annual Report and Accounts. .* Accessed 7 29, 19. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/728071/STA_annual_report_and_accounts_2017_to_2018.pdf .

Stiggins, R. J. 1991. "Assessment literacy." *Phi Delta Kappan,* 72: 534-539.

Tengburg, M. 2017. "National reading tests in Denmark, Norway, and Sweden: A comparison of construct definitions, cognitive targets, and response formats." *Language Testing* 34 (1): 83 –100.

Tennent, W. 2015 . *Understanding reading comprehension: Processes and practices. .* London: Sage.

Tennent, W., Stainthorp, R. and Stuart, M. 2008. "Assessing reading at Key Stage 2: SATs as measures of children's inferential abilities." *British Educational Research Journal, 34 (4):* 431–46.

Wiliam, D. 2010. "Standardized Testing and School Accountability." *Educational Psychologist,* 45 (2) 107-122.

Wixson, K. K. 2017. "An Interactive View of Reading Comprehension: Implications for Assessment. ." *Language, Speech, and Hearing Services in Schools. 48:* 77–83 .

Zwaan, R.A. 1996. "Toward a model of literary comprehension". In *Models of Understanding Text* by B. Britton & A.C. Graesser (eds). Mahwah, NJ: Laurence Earlbaum Associates.