UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

How to assess children's virtue literacy: methodological lessons learnt from the Knightly Virtues programme

Davison, Ian; Harrison, Tom; Hayes, Daniel; Higgins, Jenny

DOI:
10.1080/13617672.2016.1141526

License:
None: All rights reserved

Document Version
Peer reviewed version

Citation for published version (Harvard):
Davison, I, Harrison, T, Hayes, D & Higgins, J 2016, 'How to assess children's virtue literacy: methodological lessons learnt from the Knightly Virtues programme', *Journal of Beliefs and Values*.
https://doi.org/10.1080/13617672.2016.1141526

Link to publication on Research at Birmingham portal

Publisher Rights Statement:
This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of Beliefs and Values on 8th March 2016, available online: http://www.tandfonline.com/10.1080/13617672.2016.1141526

Checked March 2016

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 13. Dec. 2024

How to assess children's virtue literacy: methodological lessons learnt from the Knightly Virtues programme

Authors

Ian Davison, School of Education, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, i.w.davison@bham.ac.uk 0121 414 4808

Tom Harrison, Jubilee Centre for Character and Virtues, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, t.j.harrison@bham.ac.uk 0121 414 4812

Dan Hayes, School of Education, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, 0121 414 4855 D.S.Hayes@bham.ac.uk

Jenny Higgins, Jubilee Centre for Character and Virtues, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom, j.higgins@bham.ac.uk 0121 414 4803

Ian Davison is the corresponding author.

Notes on contributors

Ian completed his PhD in psychology in 1987; then taught science in secondary schools in various roles. Since 2004, he has worked at the University of Birmingham and been involved in 28 externally funded research projects, principally for data analysis.

Tom is the Director of Development at the Jubilee Centre for Character and Virtues at the University of Birmingham. In this role he is responsible for running national Character Education research and development interventions. He completed his PhD in education in 2014.

Dan worked for the Jubilee Centre between 2012 and 2015, providing qualitative research input for the Development Team. He is currently working on a PhD in Education Policy.

Jenny started working at the Jubilee Centre in 2012 and until May 2015 was the Development Officer, managing community outreach and projects in schools. She previously worked for the charity Volunteering Matters, managing volunteer-in-schools projects.

How to assess children's virtue literacy: methodological lessons learnt from the Knightly Virtues programme

Abstract

Character education is of growing importance in educational discourse. The Knightly Virtues programme draws on selected classic stories to teach eight moral virtues to 9-11 year olds; it has proved to be hugely popular with UK schools. A finding of the trial was the different levels of 'virtue literacy' in faith and non-faith schools (Arthur, Harrison, and Davison 2015). This paper outlines the key features of this trial which yielded the positive results and details its methodological strengths and potential shortcomings. Overall, statistical concerns are less problematic than the practical concerns of running trials designed to measure the impact of character education interventions. Of greatest theoretical interest is the conflation of general and virtue-specific literacy; in addition, we tease apart differences in understanding and motivation. The paper highlights and discusses challenges of running trials designed to measure character education as well as providing insights into promising methodological approaches.

Keywords: character; education; virtue literacy; programme evaluation; educational trial

Introduction

The revival of character education, with an emphasis on teaching moral virtues, has been a striking global educational trend over the last decade. Programmes with the aim of developing character virtues such as compassion, courage and gratitude in young people are being formally adopted into school curricula in many countries, including in the U.K. In 2014, the Department of Education announced its intention to make England 'a global leader of teaching character' and committed significant funding for the development and evaluation of character education programmes (Department for Education 2014).

It is against this backdrop that the Knightly Virtues programme was developed by the Jubilee Centre for Character and Virtues at the University of Birmingham. The programme draws on

selected classic stories to teach 9-11 year olds about eight moral character virtues (Carr and Harrison 2015). Exploring four different stories, the pupils discuss the virtues of the central characters and then relate these to their own lives. The programme sought to develop the 'virtue literacy' of the pupils, which for present purposes refers exclusively to the knowledge, understanding and satisfactory application of virtue terms, as distinct from the development of virtuous emotions or virtuous behaviours, although theories of character development provide some hope that the former may contribute to the latter (Kristjánsson 2015). In the three years since its development, over 30,000 9-11 year olds have taken part in the Knightly Virtues programme discussed in this article. The popularity of the programme, along with the professional judgement of the teachers who have elected to integrate it into their curriculum, suggests it is educationally worthwhile. However, it is not enough to assume that just because the programme is popular it is effective at fulfilling its aim of developing virtue literacy: rigorous evaluation is required.

The impact of the programme was evaluated using a before and after controlled trial. As described below, the research empirically supports the claim that these stories of Knightly Virtues helped to develop moral character. One feature of the trial was to compare pupils attending English faith and non-faith schools. Interestingly, children at Catholic schools had higher pre-test scores than those from non-faith and Church of England schools; this suggests that their grasp of virtue language and concepts was initially better developed (Arthur, Harrison, and Davison 2015). This paper reports on the main findings from the Knightly Virtues' trial and considers in detail its methodological strengths and potential shortcomings.

Conducting Trials in Character Education

Berkowitz and Bier (2005) provide a useful overview of scientific research in America on

character education to help practitioners make judgements on ‘what works’. Their paper draws together the recommendations from 33 programs which they argue have sufficient scientific backing to demonstrate their effectiveness. However, the paper says very little about the research methodology, tools and strategies employed for undertaking the programme evaluations. Elsewhere in the social sciences, trials, and in particular Randomised Control Trials (RCTs), are often seen as the gold standard to demonstrate effectiveness (Torgerson and Torgerson 2008). It is argued that such trials should form the basis for evidence-based educational policies. However, perhaps due to the difficulties of successful implementation in schools, there have been limited attempts in Britain to apply RCTs to research into character and virtue. One example is the RCT feasibility study of the My Character programme, which was designed to enhance the character virtue of future mindedness (Arthur, Harrison, Kristjansson, et al. 2014). This report demonstrated the feasibility of running RCTs in character education, whilst highlighting the difficulty in creating suitable outcome measures. In both Britain and America, few trials of a similar nature have been conducted, although the Education Endowment Fund (educationendowmentfoundation.org.uk) has recently committed to fund a series of trials in Britain that attempt to evaluate the effectiveness of character development programmes.

With a likely increase in the use of trials to measure character education interventions, this paper aims to aid their implementation by discussing the methodological strengths as well as the more problematic lessons learnt from the Knightly Virtues trial. Through this process it is envisaged that a richer understanding of both the complexity of such trials as well as recommendations for similar ventures will emerge. Robust evidence will also help to promote character education with both policy makers and practitioners.

In this paper we discuss the effectiveness of our approach to assessing how well the Knightly

Virtues programme develops virtue literacy in 9-11 year olds. An important component of virtue literacy is the acquisition of virtue language through familiarity and use of virtue terms. Although knowledge of language alone is not sufficient for virtuous behavior, the acquisition of virtue terms and concepts contribute to one's ability to undertake rational reflection and deliberation (Carr and Harrison 2015).

Putative links between virtue literacy and virtue practice have been made in the literature. Lickona (1992) proposed that an important step of character education was to develop understanding of virtue meanings in order to realise their importance. Vasalou (2012) argues that there is a link between the mastery of language and the mastery of virtue on similar grounds. Likewise, the so called 'four component model of moral behaviour' (Rest 1986) highlights the importance of virtue knowledge and understanding. This model considers reasons for moral behavior and the required capacities for successful moral functioning. The four components in the model are moral sensitivity, moral judgement, moral motivation and moral character. Virtue literacy is most closely related to moral sensitivity which focuses on the ability to identify and discern problematic situations with ethical dimensions, and moral judgement which focuses on the ability to conceptualise situations in moral language and pass satisfactory evaluative judgements on them. However, Narvaez (2002), who was also involved with the development of the four-component model, challenges assumptions that children build moral literacy from reading or hearing moral stories.

The challenge for the research team was to develop robust, valid research procedures to satisfactorily assess virtue literacy. There is little consensus on how character virtues can be measured, or indeed if they even should be measured (Kristjánsson 2015). Serious challenges have been identified when attempts are made to measure character virtues for the purposes of educational policy and practice (Duckworth and Yeager 2015). All measures currently

available have limitations. Self-report surveys are popular for assessing character; for example, see Peterson and Seligman (2004). However their validity is often challenged due to the subjectivity of self-reporting character strengths and weaknesses, and issues such as social desirability and self-delusion mean that participants may not provide unbiased accounts of their own virtues. Reports by others, such as teachers, suffer from reference bias when comparing different schools (Duckworth and Yeager 2015). An alternative is to use moral dilemmas e.g. Thoma (2006); this approach has been popular since the work of Kohlberg (1972). One of the main critiques of dilemmas is that they show what a person might theoretically do – but not necessary how they would respond in a 'real life' situation. On a practical note, which was of paramount importance for us, they are difficult to implement and analyse, particularly with young children. Attempts have been made to measure proxy outcomes of character education such as behaviour and knife crime with some success: see overview paper by (Berkowitz and Bier 2005). However, behavioural outcome measures are difficult and time-consuming, often requiring sequential assessment of individuals; whereas we were seeking an approach that could be used quickly and simply by classroom teachers. Moreover, from the theoretical perspective of character education, someone might develop a cunning capacity to engage in virtuous looking behaviour although she did not possess the relevant virtue. Fallona (2000) has successfully assessed virtue through observation; however practical issues often make them untenable. It is with these concerns in mind that the research strategy for the Knightly Virtues trial was developed. The aim was to create a valid assessment tool that assesses virtue literacy but does not suffer from overly challenging logistical or implementation concerns.

After a brief overview of the trial method and results, this paper focuses on a discussion of the lessons learnt. Here we investigate several methodological concerns in order to recommend improved approaches to trials of this nature and highlight the need for deeper

empirically-based theorizing of the nature of virtue literacy.

Method

Here we report on the 1089 pupils in 47 classes in 19 schools that completed both the pre- and post-tests.

The Knightly Virtues programme was designed to test the hypothesis that primary schools could employ classic stories to facilitate learning about character virtues. This traditional approach goes back to Aristotle who regarded the exposure to narratives as important to the education of the desires and the emotions. MacIntyre (1981) argues that narratives help us see ourselves as human and agents as they provide a logical form of human self-understanding. More recently Bohlin (2005) and Carr and Harrison (2015) have provided practical examples of ways that teachers can help pupils understand the ethical themes and issues of both the classic stories and their own lives. The original incarnation of the Knightly Virtues programme contained four stories: Gareth and Lynette from the King Arthur legends; El Cid; Don Quixote; and the Merchant of Venice. The belief was that these stories would be an attractive source for the consideration of the following virtues: gratitude, courage, humility, service, justice, honesty, love and self-discipline. All the schools that took part were provided a teaching pack consisting of lesson plans, presentations and resources and the pupils were all given a personal journal containing the stories and activities relating to the programme.

In each school, one or more classes undertook the Knightly Virtues programme and one or more classes acted as a control. If the experimental group was in Year 6 (ages 10 and 11), then the control group was in Year 5 (ages 9 and 10), and vice versa. As described below, a pre-test was given to both groups before the programme, and a similar post-test afterwards.

Six Teacher Assessors, experienced in assessing Year 5 and 6 pupils, were enlisted to help design the test and to mark the pre- and post-tests.

As the Knightly Virtues programme is based around comprehension activities and would most likely be implemented during literacy lessons, we used the Key Stage 2 (KS2) National Curriculum (NC) assessment in reading as the template for constructing the outcome measure. This meant that the pupils would be familiar with the test format, it would provide a useful literacy exercise that the teachers might value, and, most importantly, would be assessing virtue literacy as opposed to a self-reported measure.

The KS2 Reading assessments are comprehension exercises consisting of a reading booklet containing text extracts (1,800 – 2,300 words) and a question and answer booklet containing 40 – 50 questions; see Standards and Testing Agency (2014). These tests are designed to last 60 minutes and test the pupil's ability to both retrieve information from the text as well as deductive reasoning at the higher levels. The Knightly Virtues outcome measure was considerably shorter, containing approximately 1000 words split across two stories, one historical and one contemporary. The stories, questions and format were piloted and seemed to be both age appropriate and effective in engaging participants. Two versions of the test were designed to be of equal style, length and difficulty. Roughly half the pupils completed Version A before the Knightly Virtues' programme (pre-test) and Version B after the programme (post-test); vice versa for the remaining pupils.

As the Knightly Virtues programme is intended to improve virtue literacy, a mark scheme was developed for each of the following Domains:

- Reading and writing comprehension as a control variable (Domain A)
- Pupils' knowledge and understanding of virtue language (Domain B)

- Pupils' application of virtue concepts in modern day stories (Domain C)
- Pupils' application of virtue concepts in historical stories (Domain D)
- Pupils' application of virtue concepts in personal, social and cultural contexts (Domain E)
- The national curriculum reading level and sub-level of each script e.g. 3c, 4a.

Domain A and the national curriculum reading level were to assess overall reading comprehension, irrespective of virtue literacy, to act as control variables. Domain B looked at the use of virtue language beyond that explicitly referred to in the question paper. Domains C, D and E mapped to sections of the question paper i.e. C referred to the contemporary story, D to the historical story and E to additional questions relating to the pupils, themselves, and other people alive today.

Marking was on a 0 to 6 scale, using the following adjectives for each Domain: 0= 'no', 2= 'weak/ insecure', 4= 'moderate', and 6= 'strong'; e.g. 4 in Domain C means 'moderate application of virtue concepts in modern day stories'. Through discussions as well as group and individual moderated tasks, the Teacher Assessors developed a group understanding of how to interpret these descriptors.

Repeated measures analysis of variance (ANOVA) was used to investigate whether the Knightly Virtues programme improved test scores relative to controls. Therefore for each Domain separately, a Group (experimental: control) by Time (pre- and post-test) by Version (A or B) by Year (5 or 6) ANOVA was undertaken with the pre-test reading level as a covariate. For inter-rater reliability (IRR), the intra-class correlation coefficient was used with average measures, absolute agreement, assessors treated as a fixed effect and scripts as a

random effect. The principle component method of exploratory factor analysis was used with varimax rotation. These analyses were undertaken in SPSS version 21. To explore school-level effects, multilevel modelling was undertaken in MLwiN (Rasbash et al. 2009).

Research design

A before and after controlled trial design was used (see Table 1).

Table 1 about here

As described above, RCTs are usually considered the gold standard for investigating effectiveness, at least from a quantitative standpoint. However, we decided not to attempt to randomise for two reasons. First, the Knightly Virtues programme is designed to be used with the whole class. Therefore it is not possible to randomise pupils within their classes. It is, of course, possible to randomise schools, but then half the schools have to accept that they will not be undertaking the Knightly Virtues programme, at least until the end of the trial. Our judgment was that it would have been extremely difficult to recruit and retain sufficient schools in the control group. Randomising schools may create greater variance between the two groups, although the multilevel analysis suggests this between-school difference is not large. Part of our decision to include both an experimental and control class from each school was to reduce this variance (Campbell, Elbourne, and Altman 2004). In the ideal world, each participating school would have offered 2 classes that were then randomly allocated to the two groups; unfortunately, again, we judged this to be infeasible as usually the school was very clear which class they wanted to engage with the Knightly Virtues programme. The second reason not to undertake an RCT is that empirically randomised and non-randomised

trials give similar results in both medicine and medical education (Cook 2012). In several areas of medicine, there is little difference between RCT and observational studies (Vandenbroucke 2011).

Results

With all domains, the experimental group improved their overall mean score more than the control group, when controlling for the assessed national curriculum level of the scripts. This trend was not significant for Domains B and D, just short of statistical significance for Domains A ($p=0.08$) and C ($p=0.09$), and highly significant for Domain E ($p<0.001$). The face-value interpretation of these findings is that the Knightly Virtues programme may have improved pupils' scores on all domains, but this effect only reached statistical significance with Domain E i.e. the greatest (and perhaps only) impact is in learning to apply virtue concepts in personal contexts not linked to the stories in the reading booklet. More details of these findings are reported elsewhere (Arthur, Harrison, Carr, et al. 2014).

The concern that pupil scores are not independent was addressed using multilevel modelling in which pupils are nested within schools. The pattern of results using this procedure was similar to that obtained when clustering was ignored. Therefore, clustering does not appear to account for the significance of the improvement due to the Knightly Virtues programme.

Inter-rater reliability (IRR) was assessed twice. During the moderation process, IRR varied between 0.65 for Domain C and 0.86 for Reading Level. During the main marking phase, the IRR of a sample of scripts dropped to between 0.48 for Domain C post-test and 0.76 for Reading Level post-test. Although these are non-significant differences, the higher initial IRRs could be due to assessors initially working together or being more careful. It has been suggested that '0.70 would be sufficient for a measure used for research purposes' (Graham,

Milanowski, and Miller 2012, 9 p9); for the main marking, the average IRR was 0.62, which is lower than desirable. Higher IRR would reduce the error variance and so increase the likelihood of significant findings. However, with 1089 pupils in the study, we do not judge this to be a serious concern.

The rest of the paper considers methodological strengths of the methods employed, the issues that may challenge our conclusion, and implications for future evaluative research of character education programmes.

Conflation of virtue literacy with general literacy

As indicated above, a NC reading level and sub-level were assigned to each script to act as a co-variate as it is the biggest predictor of marks awarded in each Domain. A reading level from other work for each pupil would have been desirable, but many schools did not provide that information and those that did only gave the level, not sub-level, which was too crude for our analysis. We hypothesised that the NC reading level would correlate very highly with Domain A (reading and writing comprehension) as they are very similar, but we hoped that Domains B, C, D and E would be reasonably distinct from NC reading level as they were concerned with virtue language and concepts which may have largely been learnt during the Knightly Virtues programme.

Therefore, factor analysis was undertaken with two factors extracted separately for pre- and post-test scores (see Table 2 below). This table indicates that Reading Level and Domains A, B and C load heavily upon the first factor, which we have called ‘General reading comprehension’. Only Domain E (and to a lesser extent Domain D) loads heavily on the second factor, which we have called ‘Specific virtues comprehension’. This suggests that any improvement independent of Reading Level is most likely to be demonstrated with

Domain E. As expected, using the same ANOVA design as above, the Time by Group interaction is not significant for General reading comprehension (Factor 1, p=0.5), but is for Specific virtues comprehension (Factor 2, p=0.007).

Table 2 about here

This factor analysis suggests that our assessment of general reading level and Domains B and C are conflated, reducing the chance of detecting significant improvements in these domains, if indeed they are distinct from general reading skills.

Understanding or motivation?

It is difficult to distinguish between poor answers due to lack of understanding and due to lack of motivation. A crude approach is to investigate the number of times pupils do not write anything in response to a question. Therefore the number of questions not attempted was counted in a random sample of 225 papers for the three sections, which correspond to Domains C, D and E. 'Not attempted' was defined as not writing at least a word as an answer. Table 3 shows that a high percentage (87%) of pupils attempted all questions for Domain C, but this fell to 53% for Domain E. Most noticeable is that 25% of pupils did not attempt any questions in this domain. This leads to a potential explanation for the significant difference in Domain E: perhaps it is simply that more pupils in the Knightly Virtues experimental group attempted questions related to Domain E, post-test.

A Group (Knightly Virtues: control group) by Time (pre- and post-test) ANOVA was undertaken on the number of questions not attempted. The Knightly Virtues group had

significantly fewer questions not attempted for Domains D ($F(1)=15.0$), $p<0.001$) and E ($F(1) = 31.4$, $p<0.001$), and a non-significant trend for Domain C ($F(1) = 3.2$, $p=0.08$). Neither the effect of Time nor the interaction approached significance. This suggests that more effort was made in the Knightly Virtues group than control group. As this was the case during pre-tests, before the pupils had experienced the programme, it may be that the Knightly Virtues teachers were more enthusiastic or allowed more time. However, the Time by Group interaction was not significant, so this issue does not explain the significant finding that the Knightly Virtues programme improved pupils' scores in Domain E.

Table 3 about here

Other issues

Having both the experimental and control groups in the same school means there is potential for contamination between groups. Informal feedback from teachers suggests this was indeed an issue because the Knightly Virtues project moved beyond the classroom; the virtues, stories and pupil work were the subject of assemblies, wall displays and class presentations. In one case, the school adopted the project virtues as their official school virtues, disseminating the project contents to all pupils and teachers in the school. Pupils in the experimental group may have discussed work with friends in the control groups and may have shared their work with them. Consequently, pupils in control groups were exposed to Knightly Virtues work at least to some extent, which may have reduced the post-test differences between the groups.

The materials for schools contained guidance notes for delivering the project, including

estimated times, resources and structure of the lessons. However, different schools delivered the project in different ways to meet their specific needs. Consequently, there was a variety of delivery; some schools delivered the project in 90 minute lessons over 5 weeks, while others delivered it in shorter sessions over a longer period of time. In different schools, Knightly Virtues lessons were linked to literacy, history or more general project work.

From interviews, it was clear that teachers emphasised parts of the project depending on the perceived needs and interests of their pupils. One school focused on the virtue of humility as the teacher felt that this was particularly useful, while another school concentrated on the Don Quixote story, claiming it captured the pupils' interest more than the other texts.

Similar variations were found in the administering of the pre- and post-test questionnaires. Many schools completed the post-test questionnaires immediately after finishing the project, whereas with other schools there was a delay of four or five weeks. We could not discern a difference in delays between the control and experimental groups, so are not aware of an impact on our findings.

Above, we raised the possibility that teachers in the control group gave less enthusiastic instructions or less time for the pre- and post-tests. We could have given stronger guidance on administering the tests and asked for more precise recording of the time allowed for these tests.

Discussion and recommendations

The 'key finding' is that pupils undertaking the Knightly Virtues programme increased their knowledge, understanding and application of virtue language compared with controls; this effect was only significant, however, for the 'Application of virtue concepts in personal contexts' (Domain E). Elsewhere we have reported that this trial showed that 'Children

attending Catholic schools had significantly higher scores in the trials pre-test indicating that they had a better developed initial grasp of virtue language and concepts' (Arthur, Harrison, and Davison 2015, 178). Important positive features of the trial include the use of a before and after controlled trial design, 2 parallel versions of a reading test, and use of Teacher Assessors to mark the pupils' work. However, this paper has focussed on methodological challenges and consideration of potentially spurious explanations for the 'key finding'.

The factor analysis shown in Table 2 suggests that assessment of Domains B and C are conflated with the assessment of reading level. Interestingly this suggests a theoretically fruitful avenue for research. A new hypothesis is that domains B, C and to a lesser extent D draw heavily upon general reading skills as they are about interpretation of the given stories. Only Domain E requires pupils to rely upon what they have learnt in the Knightly Virtues programme. An alternative hypothesis is that it is extremely difficult to use a single reading task to assess both 'General reading comprehension' and 'Specific virtues comprehension'. To disentangle these alternative hypotheses, collection of accurate national curriculum level and sub-level independently from the schools taking part in the Knightly Virtues programme would have been extremely beneficial. This could be the class teacher's assessment, or better still, a separate task that is independently assessed to ensure consistency across classes. A separate reading task of a similar style but unrelated to character education would provide an excellent assessment of pupils' general reading level. Perhaps utopian, it may be possible to run two parallel trials with each class acting as the other's control; this would require both trials to be assessed using similarly structured literacy-based tasks in completely different areas.

In the sample investigated, 25% of pupils did not even attempt to answer any of the four questions in Domain E; this percentage was greater in the control group. However, there was

no evidence that this affected the significance of our findings. Possible reasons for the low response to Domain E include: the teacher didn't allow enough time; the pupil gave up and so didn't attempt the questions; and, the questions were too difficult (as they were not directly related to the text). Further research would be needed to disentangle these differences. However, the low average marks across all Domains suggest the tasks could be made a little easier; enabling almost all pupils to answer all questions would improve the sensitivity of the study and provide a more positive learning experience. With hindsight, we could have been much more prescriptive on how long the teacher should give the class to complete the test and to record how many minutes were allocated to the task.

Teachers adapted and extended the programme to fit their needs, so differences between the control and experimental groups may have been diminished by whole school engagement with Knightly Virtues. Alternatively, any positive findings may be due to extension activities as well as the intended programme. For a robust evaluation, the pre- and post-test procedures need to be followed as much as pragmatically possible; perhaps more important is accurate recording of what actually happened. For example, it was not possible to insist on a set length of time for the pre- and post-tests due to different timetabling constraints, but a more accurate record of the times devoted to them would have been beneficial.

In the Knightly Virtues evaluation, statistical concerns relating to clustering of data, lack of randomisation and inter-rater reliability do not seem to be serious threats to validity. More problematic are issues of trial design related to possibilities of contamination and differences in the way the pre- and post-tests were administered. It is easy to say that pupils in the control group need to be in separate schools and there must be tight adherence to protocols. However, our view is that this purist approach was simply infeasible despite strong positive relationships with participating schools. The teachers' priority is to maximise pupil

engagement and learning, and they were enthusiastic about the Knightly Virtues programme. However, most of them did not really understand the purpose of the pre-and post-tests. Increased understanding of the nature of research by participating teachers seems to be the most fruitful way forwards, which may necessitate the development of teachers as researchers.

From a theoretical perspective, investigating the distinctions between general and virtue literacy and between understanding and motivation have been most fruitful. Thus, what started as a focussed evaluation of a character education intervention has turned into a quest to understand the nature of virtue literacy and to optimise the outcome measure.

In light of the findings of this paper, we make the following recommendations/suggestions to researchers attempting evaluations of character-education programmes along similar lines to those of the Knightly Virtues programme:

A) Before embarking on a trial of any new or established character education intervention, the construct, and hence the outcome measure, requires careful consideration.

The Knightly Virtues programme seeks to improve virtue literacy. Therefore, the trial made no attempt to assess virtuous behaviour; the outcome measure was designed as a written test to determine changes in pupils' knowledge, understanding and application of virtue terms and concepts. This test was made more realistic and meaningful to pupils by modelling it upon the familiar National Curriculum reading test. Pragmatism, on the part of the research team, was required given the considerable challenges of assessing character.

B) Reading and writing comprehension must be controlled for in written tests.

As demonstrated by factor analysis, general and virtue literacy are confounded. Therefore

general reading ability should be used as a control variable. In this study, this was done by Teacher Assessors giving a global NC reading score for each pupil's pre-test answer booklet. Although this was adequate, use of a text unrelated to character education would be ideal.

C) Teacher engagement is crucial, but it is unrealistic to expect teachers to follow strict instructions; more realistic is careful recording of what they do.

Teachers are educators, not researchers, and so give children activities that they believe are rewarding. Although the Knightly Virtues research team worked closely with teachers as 'researchers in situ' to help implement the methods and tools, it cannot be claimed that the all the instructions were followed. The teachers, generally adhered to the trial procedures, but sometimes adapted them with their pupils' best interests at heart. It would be wrong to prevent this happening and researchers need to work with the grain of teaching expertise. Therefore, extensive engagement with teachers so they understand the purpose of the research and record what they do, even if slightly different from the agreed protocol, is probably the right balance between teaching and research.

The present discussion outlines the depth of thinking as well as research activity that was required to undertake an acceptable programme evaluation into character education. Duckworth and Yeager (2015) envisaged the use of online activities for this purpose. As our focus is on virtue literacy as opposed to behaviour, we believe the paper-based test format employed here is preferable. We have explored several threats to validity; whilst these threats have not undermined our finding that the Knightly Virtues programme improves application of virtue concepts in personal contexts, they point to further methodological improvements along the long road to adequate evaluation of interventions in character education.

Our analyses suggest that despite the challenges, it is possible to gain useful and significant

results from trials into the impact of new character education interventions, leading to greater understanding of what works in the classroom. The phenomenal interest in the newly expanded Knightly Virtues programmes provides confidence that improved evaluation of such programmes is a worthy educational pursuit.

Acknowledgements

We would like to thank all the young students, teachers and head teachers who made this research possible and positive. This work was supported by the Jubilee Centre for Character and Virtues with funding from the John Templeton Foundation.

References

Arthur, J., T. Harrison, and I. Davison. 2015. "Levels of Virtue Literacy in Catholic, Church of England and non-faith Schools in England: a research report." *International Studies in Catholic Education* 7 (2):178-200.

Arthur, James, Tom Harrison, David Carr, Kristján Kristjánsson, and Ian Davison. 2014. Knightly Virtues: Enhancing Virtue Literacy Through Stories: Research Report. Birmingham. ISBN: 978-0-7044-2844-7: The Jubilee Centre for Character and Virtues.

Arthur, James, Tom Harrison, Kristjan Kristjansson, and Ian Davison. 2014. My Character, Enhancing Future-mindedness in Young People: A Feasibility Study. University of Birmingham ISBN: 987-0-7044-2842-3.

Berkowitz, Marvin W., and Melinda C. Bier. 2005. What Works In Character Education: A research-driven guide for educators. Washington, DC: Character Education Partnership.

Bohlin, K. . 2005. *Teaching Character Education through Literature*. london and new york: routledgeFalmer.

Campbell, Marion K, Diana R Elbourne, and Douglas G Altman. 2004. "CONSORT statement: extension to cluster randomised trials." *British Medical Journal* 328:702-8.

Carr, D., and T. Harrison. 2015. *Educating Character through Stories*. Exeter: Imprint Academic.

Cook, D. A. 2012. "Randomized controlled trials and meta-analysis in medical education: What role do they play?" *Medical teacher* 34 (6):468-73. doi: 10.3109/0142159X.2012.671978.

Department for Education. 2014. "England to become a global leader of teaching character." Accessed 14/10/2015]. https://www.gov.uk/government/news/england-to-become-a-global-leader-of-teaching-character

Duckworth, A. L., and D. S. Yeager. 2015. "Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes." *Educational Researcher* 44 (4):237-251. doi: 10.3102/0013189x15584327.

Fallona, C. . 2000. "Manner in teaching: A study in observing and interpreting teachers' moral virtues." *Teaching and Teacher Education* 16 (7):681-695.

Graham, Matthew, Anthony Milanowski, and Jackson Miller. 2012. Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings Center for Educator Compensation Reform.

Kohlberg, L. . 1972. "A cognitive-developmental approach to moral education." In *Collected Papers on Moral Development and Moral Education*, edited by L. Kohlberg, pp. 13-16.

Kristjánsson, K. 2015. *Aristotelian Character Education*. Abingdon: Routledge.

Lickona, T. 1992. *Educating for Character: How Our Schools Can Teach Respect and Responsibility*: Bantam.

MacIntyre, A.C. 1981. *After Virtue*. Notre Dame: Notre Dame Press.

Narvaez, D. 2002. "Does reading moral stories build character?" *Educational Psychology Review* 14 (2):155-171.

Peterson, C., and M. E. P. Seligman. 2004. *Character strengths and virtues: A handbook and classification*. Oxford: Oxford University Press.

Rasbash, Jon, Fiona Steele, William Browne, and Harvey Goldstein. 2009. *A User's Guide to MLwiN: Version 2.10*: Centre for Multilevel Modelling, University of Bristol.

Rest, J. 1986. *Moral Development: Advances in Research and Theory*. New York: Prager.

Standards and Testing Agency. 2014. "Key stage 2 tests: 2014 levels 3-5 English reading test materials." Accessed 14/10/2015]. https://www.gov.uk/government/publications/key-stage-2-tests-2014-levels-3-5-english-reading-test-materials.

Thoma, S. 2006. "Research on Defining Issues Test." In *Handbook of Moral Development*, edited by M. Killen and J. G. Smetana. Mahwah, New Jersey: Erlbaum.

Torgerson, David J. , and Carole J. Torgerson. 2008. *Designing randomized trials in health, education and the social sciences: an introduction*. Basingstoke: Palgrave Macmillan.

Vandenbroucke, Jan P. 2011. "Why do the results of randomised and observational studies differ?" *British Medical Journal* 343:d7020.

Vasalou, S. 2012. "Educating Virtue as a Mastery of Language." *Journal of Ethics* 16 (1):67-87.

Table 1: research design of the Knightly Virtues trial

Group name	Pre-test	Trial activity	Post-test
Experimental 1	Version A	Knightly virtues teaching	Version B
Experimental 2	Version B	Knightly virtues teaching	Version A
Control 1	Version A	normal teaching	Version B
Control 2	Version B	normal teaching	Version A

Table 2: Factor loadings for the five domains and NC reading levels

	Pre-test factor loadings		Post-test factor loadings	
Domain	General reading comprehension	Specific virtues comprehension	General reading comprehension	Specific virtues comprehension
A	.850	.374	.782	.496
B	.786	.464	.716	.565
C	.857	.257	.812	.376
D	.540	.735	.597	.680
E	.275	.925	.303	.918
NC Reading level	.807	.339	.867	.263

Table 3: number of questions not attempted

Questions not attempted	Domain C/6		Domain D/5		Domain E/4	
	n	%	n	%	n	%
0	196	87.1	157	69.8	120	53.3
1	18	8.0	26	11.6	25	11.1
2	2	.9	18	8.0	14	6.2
3	4	1.8	8	3.6	10	4.4
4	3	1.3	6	2.7	56	24.9
5	1	.4	10	4.4	-	-
6	1	.4	-	-	-	-
Total	225	100.0	225	100.0	225	100.0
	p value		p value		p value	
Time	.344		.156		.394	
Group	.076		<.001		<.001	
Time * Group	.178		.186		.275	