Measuring self-regulation in preschool children: developing measures for studying the effects on preschool children of a professional development programme for teachers

Kathy Sylva, May Shakespeare and Fiona Jelley, Department of Education, University of Oxford

June 2019

Project Delivered by Dr David Whitebread

Managed by Laura Barbour The Sutton Trust

Funded by Headley Trust and Esmée Fairbairn Foundation (September 2018-August 2019)

Trial sites: Bird in Bush nursery, Old Church nursery and Redlands Primary

Aim of the CPD

The over-arching aim of the professional development programme was to support teachers to develop ‘self-regulated abilities in the children in their early years settings.

Aim of the ‘comparing measures’ empirical study

The aim was to field-test a wide array of self-regulation measures, including executive function that might be used in future evaluations to examine the effects of the CPD on children’s development.

Procedure

1. Ethical approval was obtained from the Oxford Central University Research Ethics Committee (opt in, followed by opt out for additional ‘hot’ tests).
2. Observing the CPD sessions: Kathy Sylva and May Shakespeare attended a selection of CPD sessions to learn about the goals of the programme, and to meet the teachers.
3. A sample of 10 children (later expanded to 15) was selected from five preschool groups at: Old Church Nursery School, Redlands Primary School, and Bird in Bush Nursery.
4. After suitable pilot work in Oxford nurseries, all children were administered a package of 5 measures at Pre-Test (November 2018) to establish how feasible the measures were for children of this age who came from (comparatively) disadvantaged neighbourhoods in East London. The teachers also completed two different rating scales for each child assessing self-regulation.
5. After expanding the sample, the same measures were administered at Post-Test (May 2019) to see if children’s scores had improved.
6. Using the expanded sample at Post Test, two new measures of ‘hot’ (emotionally involved) self-regulation were added to the testing package as a deliberate comparison for the ‘cool’ tablet-based measures that were part of the original assessments at Pre-test.

Research Questions

1. Do the measures of self- regulation show change over time? In other words, do they show ‘improvement’ in children that might be attributed to the effect of the CPD programme? Because there was no control group in this small scale study, we cannot attribute increase in children’s scores to the intervention; however, it is important to learn whether these tests can detect ‘improvement’ over a short period of time, improvement that MIGHT be attributed to positive intervention effects or to the passing of time in a good nursery.
2. Are the various measures related to one another? In other words, are they measuring the same or different things?
3. How do the measures of self- regulation relate to the British Ability Scale vocabulary test?
4. How do the direct, behavioural measures of self- regulation relate to different teacher-completed questionnaires?
5. Are the ‘hot’ (involving emotions) measures of self- regulation related to the ‘cool’ ones?
6. Are the two teacher-completed rating scales for self-regulation related to each other, i.e., measuring similar aspects of children’s development?

***Assessments (**all assessments administered at pre and post-test unless specified)

Direct tests of children implemented 1:1 by a researcher in quiet area of the nursery

- Heads-Toes-Knees-Shoulders task (HTKS task similar to 'Simon Says' game)
- Go/No Go task (tablet-based test of inhibition)
- Mr Ant task (tablet-based test of working memory)
- Card Sort task (tablet-based test of cognitive flexibility)
- Less Is More task of 'hot' self -regulation (only at post-test)
- Delay of Gratification task of 'hot' self-regulation (similar to the Marshmallow test, only at post-test)
- British Ability Scales (BAS 3) Naming Vocabulary Test (At pre-test for main sample, at post-test for expanded sample)

Teacher completed questionnaires in form of rating scales

- Child Self-Regulation and Behaviour Questionnaire (CSBQ , Howard and Melhuish, 2016)
- Children's Independent Learning Development Questionnaire (CHILD, Whitebread et al, 2009)

Participants

Forty six children were tested at pre-test (24 boys), with a mean age of 45.15 months (range of 38-54 months) at the time of pre-testing. The expanded sample at post-testing was 78 (mean age 49.43, range of 40-56).

Findings

All statistical analyses were carried out on between 38 and 44 children. Missing data reflects the fact that children were sometimes absent, had to go home early or had to see a visiting specialist. Occasionally a child refused to attempt a test. However all measures have at least 38 children to compare their pre- and post-test scores on the tasks, whereas the teachers completed questionnaires at pre- and post-test for almost all the children.

1. Two of the 3 direct tablet tests (executive function skills of inhibition and working memory) are reasonably related to one another and to the direct HTKS test. The tablet-based test of cognitive flexibility (card sort) is rarely related to other measures and appears to measure something quite distinct.
2. The 'less is more' test of hot self-regulation is related to all three tablet tests and also to the HTKS and the BAS vocabulary. However, the 'hot' delay of gratification test which is similar to the Marshmallow test is not related to the direct tests or to the teacher completed questionnaires. Most importantly, the less is more test (widely considered a test of 'hot' self-regulation) is not related to the delay of gratification test, an interesting finding showing how unique is the delay of gratification task.

3. The two questionnaire measures are reasonably related to one another, meaning that they are measuring – more or less – the same thing. However, the CSBQ sub scales of 'externalising problems' and 'emotional Self-regulation' are not related at all to the CHILD. Moreover, the CHILD 'motivation' sub-scale is rarely related to the CSBQ subscales, suggesting that the motivational element in the CHILD is not picked up in most of the measures of self-regulation in the published literature.
4. All four subscales of the CHILD are related to the 'less is more' test of SR whereas only one sub scale of the CSBQ (sociability) is related to the 'less is more' test. Thus the CHILD is a better predictor of one of the 'hot' measures. Note that the delay of gratification test is rarely related to any of the measures in this study, with the exception of its correlation with the BAS vocabulary. It is also of interest that the delay of gratification test is not related to age in months, although the less is more test is modestly correlated with age. Again, the delay of gratification task is unusual and less 'cognitive' than all other tests.
5. The 'motivation' sub scale of the CHILD relates well to other subscales on the same questionnaire but not much to the direct tests. The motivation subscale on CHILD is only modestly related to the CSBQ. Thus the motivation subscale of CHILD measures something not measured by other assessments in this battery.
6. Almost all of the measures showed significant improvement between pre and post-test. The only exception to the general rule of improvement was the Card Sort tablet-based test.

To summarise, children improved (scored significantly higher) at post-test in the following measures:

- HTKS
- Go No Go (inhibition)
- Mr Ant (working memory)
- CSBQ Sociability
- CSBQ Prosocial
- CSBQ Behavioural SR
- CSBQ Cognitive SR
- CSBQ Emotional SR
- CSBQ Total
- CHILD Emotional
- CHILD Prosocial
- CHILD Cognitive
- CHILD Motivation
- CHILD Total

Children improved (scored significantly lower at post-test) on undesirable measures:

- CSBQ Externalising Problems
- CSBQ Internalising Problems

Children did not change significantly at post-test on the Card Sort tablet test (cognitive flexibility).

Implications of the findings

1. What do the findings tell us about self-regulation?

Self-regulation has a huge research literature and different authors concentrate of different aspects of the construct. Blair (2003) says that 'self- regulated behavior generally refers to controlled, cognitive monitoring of the actions and steps required to obtain a goal, or to bring about a desired response from the environment'. He adds that 'executive function' skills of working memory, cognitive flexibility and inhibition all support goal directed behaviour and can be considered part of a broader notion of self-regulation.

In this study, the tablet games were a direct measure of executive function, as was the HTKS game, and they were all reasonably correlated with one another. The teacher completed questionnaires were modestly related to the direct (researcher administered) tests, especially the 'Mr Ant' test of working memory. The four direct tests administered in the pre-test appear to be measuring the same thing, and are related to some of the sub-scales on the teacher questionnaires.

Some assessments stood out as being 'different'. Strikingly, the delay of gratification (Marshmallow) test was unrelated to most other measures, for example, it was not significantly related to the 'less is more' hot test. Another 'stand-out' is the Motivation sub-scale of the CHILD related to only one of the direct tests of self-regulation but with some significant correlations with the CSBQ. A general finding is that the two teacher-completed questionnaires were often related to one another, and the direct, researcher-administered direct tests were related to one another.

Finally, the card sort test on the i-pad was rarely related to other measures in this study. The current study shows that self-regulation is not a unitary construct and that 'capturing' it in a single test is impossible. However, the CHILD, with its four different subscales, comes closer to capturing a broad definition of self-regulation than any of the other measures. But the CHILD requires further validation and tests to establish its reliability. It would also benefit from factor analysis because the sub-scales seem to be theoretically derived and not empirically so.

2. *Which assessments measure the same or similar things? Which measure something unique?*

The CSBQ and the CHILD are measuring more or less the same things. (However the CHILD predicts the hot 'less is more' test whereas the CSBQ does not). Two of the three tablet tests appear to measure something similar to what is measured in HTKS. The two cool tests of self-regulation are clearly measuring different things; the hottest test of all, the Marshmallow Delay of Gratification, is unrelated to the other measures and appears to be picking up something that no other measure are sensitive to. Finally, the 'Mr Ant' tablet test of working memory also stands out as different, although it is correlated with BAS vocabulary.

3. How might these tests be used to evaluate interventions aimed at improving self-regulation in preschool children?

Settings and schools wishing to evaluate children’s progress in response to an intervention will do well to use the CHILD, which has been validated in this study through its close association with the CSBQ (which has been validated extensively against other instruments). However, any formal evaluation of the Whitebread professional development programme could not be evaluated via the CHILD since this rating scale is part of the programme itself. Showing improvement on the CHILD would only demonstrate success at ‘teaching to the test’. However, formal evaluation of the Whitebread programme could use the CSBQ questionnaire which is an independent measure and appears to measure the same things as the CHILD. Because the CSBQ is only modestly related to the researcher administered direct tests of self-regulation/executive function, a good outcome battery in any future evaluation would include at least one or two of the direct child tests.

Since all the measures showed significant gains over time, it is suggested that all might be appropriate as evaluation tools to detect the effects of an intervention. In the absence of a control group, the study cannot show that the programme was successful. However, children’s gains on virtually all the measures are compatible with the view that the intervention improved children’s self-regulation. This is a plausible hypothesis but it will require a proper RCT.

Acknowledgements

The authors are grateful to the teachers in the three settings who completed the questionnaires on all the children and gave early feedback on administration of the direct tests. The design and implementation of this study benefited from insightful comments by Laura Barbour from the Sutton Trust and David Whitebread from the University of Cambridge.

References

Blair, C. (2003). Behavioral inhibition and behavioural activation in young children: relations with self-regulation and adaptation to preschool in children attending Head Start. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 42*(3), 301-311. https://doi.org/10.1002/dev.10103

Carlson, S.M., Davis, A.C., & Leach, J.G. (2005). Less is more: executive function and symbolic representation in preschool children. *Psychological Science, 18*(8), 609-616. https://doi:10.1111/j.1467-9280.2005.01583.x

Howard, S.J., & Melhuish, E. (2016). An early years toolbox for assessing early executive function, language, self-regulation, and social development: validity, reliability, preliminary norms. *Journal of Psychoeducational Assessment, 35*(3), 255-275. https://doi:10.1177/0734282916633009

Ponitz, C. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly, 23*(2), 141–158. https://doi:10.1016/j.ecresq.2007.01.004

Prencipe, A., & Zelazo, P.D. (2005). Development of affective decision making for self and other: evidence for the integration of first- and third-person perspectives. *Psychological Science, 16*(7), 501-505. https://doi:10.1111/j.0956-7976.2005.01564.x

Whitebread, D., Coltman, P., Pasternak, D.P., Sangster, C., Grau, V., Bingham, S., Almeqdad, Q., & Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition Learning, 4*(1), 63-85. https://doi:10.1007/s11409-008-9033-1

Appendix A Participants and ethics

Staff in the three early years settings were all participating in a year-long professional development programme for teachers to enhance their support for the development of self-regulation. They all agreed to participate in a study of comparing measures of self-regulation.

Forty six children were tested at pre-test (24 boys), with a mean age of 45.15 months (range of 38-54 months) at the time of pre-testing. One child left the school before post-testing started, so their data are only included in pre-test. At post-test, the same pre-test children (now *N* = 45) were re-tested, who had a mean age of 49.93 months (range of 44-56) at the time of post-testing. At the same time as post-testing, the sample was expanded to include 32 more children (16 boys) to be tested alongside the original sample. These 32 children had a mean age of 48.44 months (range of 40-56 months) at the time of post-testing. The whole sample of 78 children (40 boys) had a mean age of 49.31 (range of 40-56 months) at the time of post-testing.

The study was approved by Oxford University's Central University Research Ethics Committee (CUREC) in 2018 and 2019. The parents of all participating children gave signed consent for their children to participate in the direct tests and the teachers to complete 'rating scale profiles' for their children. Opt-out consent was given by parents for the second researcher to add the hot tests to the assessment battery for which they had previously given opt-in consent.

Appendix B: Measures and testing procedures

During pre-testing, 'cool' self-regulation was directly measured through the Heads-Toes-Knees-Shoulders task (Ponitz et al., 2008) and three of the iPad-based games available in the Early Years Toolbox (Howard & Melhuish, 2016): Go/No Go, Mr Ant, and Card Sort. Practitioners in the preschool and nursery settings completed two questionnaires to measure children's level of self-regulation and social behaviour. Language was measured using the British Ability Scales (3rd Ed.) Naming Vocabulary Test. All of these measures were repeated at post-testing, along with the addition of the Less Is More game (Carlson, Davis, & Leach, 2005) and the Delay of Gratification game (Prencipe & Zelazo, 2005) to measure 'hot' self-regulation.

The sample size was expanded at post-testing by recruiting similar children from the original classes. This was done to increase power but also to add two additional tests of 'hot' (emotionally involved) self-regulation.

Heads-Toes-Knees-Shoulders task

The Heads-Toes-Knees-Shoulders (HTKS) task was developed by Ponitz et al. (2008) as a measure of behavioural regulation. This includes the ability to stop an automatic response and demonstrate a different behaviour (inhibitory control), focusing and shifting attention, and the ability to hold relevant information in the mind (working memory) (Ponitz et al., 2008). In the game, children are told to do the opposite of the researcher's instructions, so when the researcher says to "Touch their toes", they should touch their head, and vice versa. Instructions could be repeated up to three times following two questions to check understanding and four practice trials. Ten testing trials are then administered, where the researcher does not give feedback to the child's responses. A correct response earned two points, a self-corrected response (defined as a child making any discernible notion towards the incorrect answer, but then changing their mind to the correct answer) was worth one point, and zero points were given for an incorrect response. If a child responds correctly to more than five testing trials, they continue to the next stage, where children are instructed to do the opposite of the researcher's instructions, this time using their knees and shoulders. After one question to check understanding and four practice trials, ten testing trials are administered, incorporating both sets of rules (heads-toes and knees-shoulders). Therefore, the total range of possible scores is 0-40.

Go/No Go task

The Go/No Go task is a measure of inhibitory control, developed by Howard and Melhuish (2016) as part of their Early Years Toolbox. Children are instructed to catch the fish by tapping the screen when they see a fish (the 'Go' trials) and to avoid the sharks by not tapping the screen when they see a shark (the 'No Go' trials). The majority of the stimuli are 'Go' trials (80%) in order to develop a prepotent response when they see a fish; therefore this requires the children to inhibit this response when they see a shark during 'No Go' trials, which occur 20% of the time. In the game, instructions for the 'Go' trials are given, followed by five practice 'Go' trials. Then, instructions and five practice trials for the 'No Go' trials are given. Combined 'Go'/'No Go' instructions are then given, with the ten practice trials involving a mix of 'Go' and 'No Go' trials (80% 'Go'). The game uses sound to provide feedback during the practice trials. The testing stage consists of three blocks of stimuli, where 75 test trials are divided evenly with short breaks and a recap of instructions in between each block. The order of the stimuli is pseudo-random, meaning that the order differs for each child but a block never begins with a 'No Go' trial and 'No Go' trials never appear consecutively. For each trial, the stimulus (fish or shark) is presented for 1.5 seconds, with a 1 second gap between each trial. An index of inhibitory control is calculated by multiplying the accuracy on 'Go' trials with the accuracy on 'No Go' trials; the total range of possible scores is 0-1.

Mr Ant task

The Mr Ant task is a measure of working memory, developed by Howard and Melhuish (2016) as part of their Early Years Toolbox. Children are asked to remember the location of 'stickers' on a cartoon picture of an ant ('Mr Ant') and point to these locations afterwards, also remembering the order which these stickers appeared. Instructions and three practice trials allow the child to become familiar with the rules before the testing commences. Test trials become harder as the game continues from Level 1 to Level 8. Within in each level, there are three trials where the number of stickers ranges from one to eight, corresponding to the current level. For each trial, Mr Ant is shown with stickers on his body for five seconds, followed by a blank screen for four seconds, followed by Mr Ant reappearing without any stickers. Auditory prompts then instruct the child to recall where the stickers were by pointing at the screen. The game continues until all levels are complete, or until all three trials at one level are answered incorrectly, whichever is earliest. A score of working memory is calculated using the following method: "beginning from Level 1, one point for each consecutive level in which at least two of the three trials were performed accurately, plus 1/3 of a point for all correct trials thereafter" (Howard & Melhuish, 2016). Therefore, the total range of possible scores is 0-8.

Card Sort task

The Card Sort task is a measure of cognitive flexibility, developed by Howard and Melhuish (2016) as part of their Early Years Toolbox. Children are instructed to sort cards (consisting of pictures of rabbits or boats) by different dimensions, either by colour (i.e., red or blue) or by shape (i.e., rabbit or boat). Before each trial, children are reminded of the current sorting dimension. After one demonstration and two practice trials, children are required to complete six test trials by sorting by colour (the pre-switch phase). In the next six trials, children are told to switch to the new rule of sorting by shape; this is called the post-switch phase. If children correctly sort at least five of the six cards in the post-switch phase, the game continues to a new phase where children are instructed to sort by colour if the card has a black border, or sort by shape if the card does not have a black border around it. Following a demonstration and two practice trials, children are required to complete six test trials using this new sorting dimension, where three trials involve bordered cards and three involve non-bordered cards. The order of the stimuli is such that, for all trials in all phases, a particular stimulus is never shown more than twice in a row. A score of cognitive flexibility is calculated using the total number of correct trials after the pre-switch phase; therefore the total range of possible scores is 0-12.

Child Self-Regulation and Behaviour Questionnaire (CSBQ)

The Child Self-Regulation and Behaviour Questionnaire was developed by Howard and Melhuish (2016) as a measure of self-regulation and social development as part of their Early Years Toolbox, School practitioners rated each child on 34 items using a Likert-type scale based on the general frequency of the target behaviours, 1 being 'Not True' and 5 being 'Very True'. Examples of positively-worded items include: "Chosen as a friend by others", "Is cooperative", and "Good at following instructions". Examples of negatively-worded items include: "Aggressive to children", "Shows wide mood swings", and "Will wander around aimlessly". Specific item scores were averaged to form seven subscales: Sociability, Externalising Problems, Internalising Problems, Prosocial Behaviour, Behavioural Self-Regulation, Cognitive Self-Regulation, and Emotional Self-Regulation. For five subscales, scores on negatively-worded items were reversed prior to analysis, so the higher the children score on these scales, the more they show these behaviours. For the two remaining subscales (Externalising Problems and Internalising Problems), all items were negatively-worded and were not reversed before analysis, therefore the higher the children score on these scales, the more they show Externalising and Internalising Problems.

Children's Independent Learning Development (CHILD) questionnaire

The CHILD questionnaire was developed by Whitebread et al. (2009) as a measure of metacognition and self-regulation. School practitioners rated each child on 22 items using a Likert-type scale based on whether each item was true of the child, 0 being 'Never', 1 being 'Sometimes', 2 being 'Usually', and 3 being 'Always'. All of the items were positively-worded, for example: "Is aware of own capabilities" and "Enjoys solving problems". Items were added to form four subscales representing different factors of self-regulation: Emotional, Prosocial, Cognitive, and Motivation.

British Ability Scales (BAS) 3 Naming Vocabulary Test

The BAS 3 Naming Vocabulary Test measures children's expressive language and their ability to name objects. Children are shown pictures of various objects of increasing difficulty and are asked to name them. There are 36 pictures in total and there are various stopping rules to apply, for example, if the child answers incorrectly for 5 consecutive pictures. Using the child's age and raw scores, T scores are calculated and used for analysis instead of raw scores, to account for the fact that children were tested on a different number of items due to these stopping rules.

Less Is More Task

The Less Is More task is adapted from Carlson, Davis and Leach (2005) and is a measure of 'hot' self-regulation where children must inhibit a 'hot' temptation. At the beginning, children are presented with two plates, one with more stickers on it than the other (five stickers compared to two), and are asked to point to the one they like better. Children are expected to prefer the one with more stickers. After the preference check, a dragon puppet called Chris is introduced and it is explained that, when the child points to a plate, Chris will get the stickers on that plate and the child will get the stickers from the other plate. Two practice trials with both verbal instructions and visual demonstration are given before the test. Each child then receives 16 test trials with only visual demonstration of their choices. The researcher gives no verbal feedback to the child during the test trials. A correct response receives one point, where the child points to the plate with the fewer stickers so that Chris receives this plate and they receive the plate with the more stickers. An incorrect response receives zero points, with the total range of scores being 0-16. Between trials, researcher reloads the plates with prepared sets of stickers. An equal number of trials involve the most stickers being presented on left and right plates. After eight trials, regardless of the child's performance, the puppet and its plate are moved from the left-hand side of the child to the right-hand side and a verbal reminder of the rule is given.

Delay of Gratification task

The Delay of Gratification task was adapted from Prencipe and Zelazo (2005) and measures children's ability to choose between an immediate reward of lower value and a delayed reward of higher value. Children are presented nine 16 cm x 9 cm cards. Based on the number of stickers and toys depicted on the cards, they can decide whether they want to get a small reward immediately or wait until they go home to win a large reward. Nine trials were administered by crossing three types of rewards (shiny stickers, plain stickers, dinosaur toys) and three types of choices (one now vs. two later, one now vs. four later, one now vs. six later). Two demonstration trials are presented both verbally and visually to show children the consequences of "immediate" and "delayed" choices. Each child then receives nine test trials without verbal feedback. For each type of reward, three cards with different number of delay choices are presented in random order. One point is awarded for correct responses, where children delay gratification and choose to receive the higher value reward; incorrect responses receive zero points. Therefore, the total range of possible scores is 0-9.

Procedure

Consent forms and information sheets were sent out to parents prior to the testing. These included details about the aims of the research and the tasks that the children would complete, as well as information about how parents could consent for their child to participate. If they wanted their child to participate, they were required to return the opt-in form to their child's class teacher.

Pre- and post-testing occurred approximately six months apart. The same procedure was followed at pre- and post-testing, although a second researcher carried out the hot tests at post-test. A researcher visited each setting for approximately two or three consecutive days to complete each set of testing, working one-to-one with each child for about a 20 minute period. In order to maintain maximal concentration, the tasks were completed in the same order for each child with the aim of keeping the child's interest. The researcher started with the BAS Naming Vocabulary test as an 'attractive' opening task, followed by the HTKS task which required maximum concentration, and then ended with the more playful iPad-based Early Years Toolbox (EYT) games. Prior to pre-testing, the order of the three i-pad games for each child was decided by randomising the order, and this order was replicated for each child during the post-testing in order to minimise order effects and keep the order of testing consistent across pre- and post-testing. Practitioners in the settings completed the two questionnaires during the pre- and post-testing phases and handed them to the researchers when they visited each setting.

With the exception of the two hot tests, all direct tests were administered by a research assistant (second author) from the Department of Education, University of Oxford.   The hot tests were administered by an M Sc student from the same department. The two rating scale questionnaires were completed at pre- and post-test by the teacher who knew the child best. Analyses in this report were conducted by the authors.