**Appendix S1: Comparison between Training and Control Groups in Control Measure Performance**

To ensure that the participants were within the typical range of abilities that might affect their irony comprehension and that the results would reflect normative development, we assessed their working memory capacity, level of language skill (i.e., reading comprehension and technical reading skill), and empathy skill.

*Working memory capacity* was assessed with the Digit Span subtest of WISC-IV (Wechsler, 2010). The control (*M* = 12.56, *SD* = 1.79) and training groups (*M* = 13.08, *SD* = 2.57) did not significantly differ in their working memory capacity, *t*(73) = -1.12, *p* = .265, *d* = 0.26.

*Technical reading skill* was measured using the word fluency subtest of Lukilasse II (Häyrinen et al., 2013) where children have to read correctly as many words as possible from a list of 105 words within a 120-s time limit. The control (*M* = 80.43, *SD* = 15.05) and training groups (*M* = 86.42, *SD* = 10.58) did not significantly differ in their technical reading skill, *t*(73) = -1.94*, p* = .055*, d* = 0.45.

*Reading comprehension* was measured using the Maze task (Ronimus et al., 2022), a paper-and-pencil task comprising 16 texts with 4 words missing from each text. Based on the textual cues, participants choose the missing word from four options. The Maze task was administered to the whole class at the same time. One of the training group participants was not present for Maze task testing and is missing this information but was otherwise kept in the analyses. The control (*M* = 35.29, *SD* = 7.94) and training groups (*M* = 35.73, *SD* = 10.22) were comparable on their reading comprehension, *t*(72) =-0.12*, p* = .902, *d* = 0.03.

*Empathy skill* was assessed using a Finnish translation of the Index of Empathy for Children and Adolescents (Bryant, 1982; Olkoniemi et al., 2023). The control (*M* = 14.51, *SD* = 2.40) and training groups (*M* = 13.16, *SD* = 3.51) did not significantly differ in their empathy skill, *t*(73) = 1.91*, p* = .059*, d* = 0.44.

**References**

Bryant, B. K. (1982). An index of empathy for children and adolescents. *Child Development, 53*, 413–425. [https://doi.org/10.2307/1128984](https://psycnet.apa.org/doi/10.2307/1128984)

Häyrinen, T., Serenius-Sirve, S., & Korkman, M. (2013). *Lukilasse 2: Lukemisen, kirjoittamisen ja laskemisen seulontatesti 1.–6. vuosiluokille. [Lukilasse 2: Screening test for reading, writing and calculus for 1st to 6th grades.]*. Hogrefe Kustannus Oy.

Olkoniemi, H., Halonen, S., Pexman, P. M., & Häikiö, T. (2023). Children’s processing of written irony: An eye movement study. *Cognition, 238,* 105508*.* <https://doi.org/10.1016/j.cognition.2023.105508>

Ronimus, M., Tolvanen, A., & Hautala, J. (2022). The roles of motivation and engagement in computer-based assessment of children's reading comprehension. *Learning and Individual Differences, 98*, 102197. <https://doi.org/10.1016/j.lindif.2022.102197>

Wechsler, D. (2010). WISC IV - Wechsler Intelligence Scale for Children (4th Ed.). NCS Pearson Ltd.

**Appendix S2: Material Pre-test**

The new experimental items used in the study were pretested using an online survey tool (Webropol, [www.webropol.com](http://www.webropol.com)) with separate groups of adults and children rating item naturality, target phrase familiarity, and comprehensibility.

First, a total of 28 experimental stories were written. A total of 107 Finnish-speaking adults (85 women, 16 men, 3 others, and 3 preferred not to answer) between ages of 19–64 (*MAge* = 33.55, *SDAge* = 10.83) rated the naturality of the different story versions on a scale from 0 (not natural) to 10 (very natural). To be included in the study, the mean difference of the rating between story versions needed to be < 2. Based on this criterion, 10 stories were excluded, leaving us 18 stories.

Another group of 29 Finnish-speaking adults (26 women, 1 man, and 2 preferred not to answer) between ages of 20–59 (*MAge* = 25.14, *SDAge* = 7.02) rated the familiarity of the target phrase as literal and ironic on a scale from 0 (not familiar) to 10 (very familiar). The target phrases were evaluated to be less familiar as ironic than as literal, *t(*17) = 2.67, *p* = .016, *d* = 0.51.

Last, 23 native Finnish-speaking children (11 girls; *M*Age = 9, range: 8–12) evaluated the comprehensibility of the experimental items. None of them participated in the actual experiment. They were asked the same text memory and inference questions after reading as in the actual experiment, but without eye-tracking. The new stories were within expected range of children’s irony comprehension when compared to the previous studies (see Table S1; e.g., Fuchs, 2023).

|  |  |  |  |
| --- | --- | --- | --- |
| **Table S1** |  |  |  |
| *Descriptive statistics for preratings of the materials used in the experiment.* |
| *Measure* | *Items* | *Literal* | *Irony* |
| Naturality | Olkoniemi et al. (2023) | 7.12 (1.17) | 6.63 (1.16) |
|  | New items | 6.72 (1.16) | 6.12 (1.31) |
| Familiarity | Olkoniemi et al. (2023) | 5.76 (1.37) | 2.43 (1.50) |
|  | New items | 5.40 (1.92) | 4.38 (2.12) |
| Comprehensibility | Olkoniemi et al. (2023) | 0.92 (0.27) | 0.63 (0.48) |
|   | New items | 0.96 (0.20) | 0.73 (0.44) |
| Note. Number of items in the Olkoniemi et al. (2023) study was *n* = 26, and new items *n* = 18. Familiarity and naturality were evaluated with ratings between 0–10 (higher score = more natural/familiar). Comprehensibility is the proportion of correct answers to inference question. |

**References**

Fuchs, J. (2023). 40 years of research into children’s irony comprehenesion: A review. *Pragmatics & Cognition, 30*, 1–30. <https://doi.org/10.1075/pc.22015.fuc>

Olkoniemi, H., Halonen, S., Pexman, P. M., & Häikiö, T. (2023). Children’s processing of written irony: An eye movement study. *Cognition, 238,* 105508*.* <https://doi.org/10.1016/j.cognition.2023.105508>

**Appendix S3: Analyses and the Final Models**

**Analyses**

 Fixations < 50 ms were either merged with a nearby fixation if the distance between the fixations was < 1° or removed from the data. As described in the pre-registration (<https://osf.io/z3xyc>), for the pre- and post-test reading data sentence-level measures were calculated (Hyönä et al., 2003; see Table S2 for description). The probability of children doing later rereading (i.e., look-back and look-from fixations) was near floor (range: .11 – .24), and analyzing the number of fixations would not give a reliable overall picture of reading. Thus, the probability of looking back to the target phrase and to initiate a look-from from the target phrase were analyzed instead. This deviates from the pre-registration but was employed for example by Häikiö et al. (2018) who studied sentence-level processing of 2nd grade children (i.e., 8-year-olds).

|  |
| --- |
| **Table S2** |
| *Descriptions of the Eye Movement Measures Used* |
| *Measure* | *Description* |
| First-pass reading time | Summed duration of fixations made on one sentence before moving to the next. |
|  | Forward-fixation time | Summed duration of fixations that land on unread parts of the sentence during first-pass reading. |
|  | Number of first-pass rereading fixations  | Sum of fixations made reinspecting a sentence before moving on in the text |
| Number of look-back fixations | Sum of fixations returning to the sentence from other parts of the text made after the first-pass reading |
| Number of look-from fixations  | Sum of look-back fixations that were initiated from the sentence. |

The data were analyzed with linear or generalized linear mixed-effects models (Baayen et al., 2008) using the lme4 package (Bates et al., 2015) in R statistical software (Version 4.1.2; R Core Team, 2021), except for the dependent variables with a high number of zero values (i.e., number of first-pass rereading fixations on target phrase, probability to look-back to target phrase, probability to initiate a look-from the target phrase, and probability to look-back to critical context) which were analyzed using *glmmTMB* package (Version 1.1.7.; Brooks et al., 2017). Separate models were built for each eye movement measure for the different text regions (target phrase, critical context, and spillover region), and for inference and text memory question accuracies. All the measures were analyzed for the target phrase, and first-pass reading time was analyzed for the spillover region. As the content of the critical context sentences was not tightly matched between Story Types (literal vs. ironic), it was not possible to reliably compare reading times on this region. Consequently, probability to look-back (binomial measure) was analyzed.

*Story Type* (Literal vs. Irony), *Group* (Control vs. Training), and *Time* (pre- vs. post-test) were fitted in the models as deviation coded fixed effects. In the pre-registration, *Time* was planned to be fit into the models as a treatment coded variable (pre-test phase set as baseline). However, at the time of testing, we considered deviation coding to be a better choice, as the intercept value in the models will indicate average reading time over both testing phases (i.e., pre-test and post-test) instead of reading time in pre-test (see Schad et al., 2020 for a review of contrasts). Participants and items were fitted into the models as random intercepts, and *Story Type* and *Group,* andtheir interaction,were fitted as random slopes. If a model failed to converge, it was trimmed top-down starting with removing covariance between random effects (Brauer & Curtin, 2018). The reading time measures were skewed and consequently logarithmically transformed by selecting the best fitting transformation using Box-Cox Power transform (Box & Cox, 1964).

For the models with a high number of zeros, after finding the random effect structure, the model was tested for zero-inflation using testZeroinflation function from DHARMa package (Version 0.4.6.; Hartig, 2022). Despite the high number of zero values, no model showed issues with zero-inflation and zero-inflation part was not fitted to the models. All the model summaries are reported in Tables S3 – S10 below.

|  |  |  |  |
| --- | --- | --- | --- |
| Table S3 |  |  |  |
| *Final Model for Inference Question Accuracy* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 75 | 1.32 | 1.15 |
| Item (Intercept) | 44 | 0.51 | 0.71 |
| Item (Story Type) |  | 1.46 | 1.21 |
| Item (Group) |   | 0.45 | 0.67 |
| **Fixed Effects** | *Estimate* | *95% CI* | *z* |
| Intercept | 1.61 | 1.27 – 1.95 | **9.23** |
| Story Type (Literal vs. Irony) | -3.18 | -3.56 – -2.80 | **-16.40** |
| Group (Control vs. Training) | -0.06 | -0.63 – 0.51 | -0.20 |
| Time (Pre-Test vs Post-Test) | 0.26 | 0.16 – 0.36 | **5.04** |
| Story Type ✕ Group | 1.13 | 0.91 – 1.36 | **9.90** |
| Story Type ✕ Time | 0.86 | 0.66 – 1.06 | **8.44** |
| Group ✕ Time | 0.59 | 0.39 – 0.79 | **5.66** |
| Story Type ✕ Group ✕ Time | 0.92 | 0.52 – 1.33 | **4.45** |
| *Note. z* values > |1.96| are bolded. |

|  |  |  |  |
| --- | --- | --- | --- |
| Table S4 |  |  |  |
| *Final Model for First-Pass Reading Time on Target Phrase* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 72 | 0.08 | 0.29 |
| Participant (Story Type) |  | 0.003 | 0.05 |
| Item (Intercept) | 44 | 0.08 | 0.29 |
| Item (Story Type) |  | 0.003 | 0.06 |
| Item (Group) |  | 0.002 | 0.04 |
| Item (Story Type ✕ Group) |  | 0.004 | 0.06 |
| Residual |   | 0.10 | 0.31 |
| **Fixed Effects** | *Estimate* | *95% CI* | *t* |
| Intercept | 7.32 | 7.21 – 7.43 | **129.98** |
| Story Type (Literal vs. Ironic) | 0.04 | 0.002 – 0.07 | **2.08** |
| Group (Control vs. Training) | -0.08 | -0.22 – 0.06 | -1.14 |
| Time (Pre-Test vs Post-Test) | -0.08 | -0.11 – -0.05 | **-5.62** |
| Story Type ✕ Group  | -0.02 | -0.08 – 0.05 | -0.51 |
| Story Type ✕ Time | -0.03 | -0.09 – 0.02 | -1.17 |
| Group ✕ Time | 0.01 | -0.05 – 0.06 | 0.23 |
| Story Type ✕ Group ✕ Time | -0.06 | -0.18 – 0.05 | -1.14 |
| *Note. t* values > |1.96| are bolded.  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table S5 |  |  |  |
| *Final Model for Forward-Fixation Time on Target Phrase* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 72 | 0.08 | 0.28 |
| Item (Intercept) | 44 | 0.06 | 0.24 |
| Item (Story Type) |  | 0.001 | 0.02 |
| Residual |   | 0.06 | 0.25 |
| **Fixed Effects** | *Estimate* | *95% CI* | *t* |
| Intercept | 7.10 | 7.00 – 7.20 | **144.92** |
| Story Type (Literal vs. Ironic) | 0.02 | -0.01 – 0.04 | 1.27 |
| Group (Control vs. Training) | -0.07 | -0.21 – 0.06 | -1.11 |
| Time (Pre-Test vs Post-Test) | -0.07 | -0.09 – -0.05 | **-6.31** |
| Story Type ✕ Group  | -0.04 | -0.09 – 0.005 | -1.75 |
| Story Type ✕ Time | -0.003 | -0.05 – 0.04 | -0.14 |
| Group ✕ Time | 0.01 | -0.04 – 0.05 | 0.33 |
| Story Type ✕ Group ✕ Time | -0.05 | -0.14 – 0.04 | -1.13 |
| *Note. t* values > |1.96| are bolded.  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table S6 |  |  |  |
| *Final Model for Number of First-Pass Rereading Fixations on Target Phrase* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 72 | 0.21 | 0.45 |
| Participant (Story Type) |  | 0.10 | 0.32 |
| Item (Intercept) | 44 | 0.32 | 0.56 |
| Item (Story Type) |  | 0.10 | 0.32 |
| Item (Group) |  | 0.12 | 0.34 |
| Item (Story Type ✕ Group) |   | 0.22 | 0.47 |
| **Fixed Effects** | *Estimate* | *95% CI* | *z* |
| Intercept | 0.23 | 0.03 – 0.43 | **2.23** |
| Story Type (Literal vs. Ironic) | 0.10 | -0.06 – 0.25 | 1.23 |
| Group (Control vs. Training) | -0.02 | -0.28 – 0.23 | -0.19 |
| Time (Pre-Test vs Post-Test) | -0.14 | -0.21 – -0.07 | **-3.90** |
| Story Type ✕ Group  | 0.23 | -0.05 – 0.50 | 1.63 |
| Story Type ✕ Time | -0.17 | -0.31 – -0.03 | **-2.36** |
| Group ✕ Time | -0.03 | -0.17 – 0.12 | -0.36 |
| Story Type ✕ Group ✕ Time | -0.26 | -0.54 – 0.02 | -1.80 |
| *Note. z* values > |1.96| are bolded.  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table S7 |  |  |  |
| *Final Model for the Probability to Initiate a Look-Back to the Target Phrase* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 72 | 0.64 | 0.80 |
| Participant (Story Type) |  | 0.23 | 0.48 |
| Item (Intercept) | 44 | 0.10 | 0.32 |
| Item (Story Type) |  | > 0.001 | > 0.001 |
| Item (Group) |  | > 0.001 | > 0.001 |
| Item (Story Type ✕ Group) |   | 0.15 | 0.39 |
| **Fixed Effects** | *Estimate* | *95% CI* | *z* |
| Intercept | -1.74 | -2.00 – -1.48 | **-13.07** |
| Story Type (Literal vs. Ironic) | 0.27 | -0.05 – 0.59 | 1.65 |
| Group (Control vs. Training) | 0.16 | -0.30 – 0.62 | 0.67 |
| Time (Pre-Test vs Post-Test) | -0.48 | -0.73 – -0.24 | **-3.85** |
| Story Type ✕ Group  | 0.09 | -0.49 – 0.68 | 0.32 |
| Story Type ✕ Time | -0.94 | -1.43 – -0.44 | **-3.73** |
| Group ✕ Time | 0.43 | -0.06 – 0.92 | 1.74 |
| Story Type ✕ Group ✕ Time | -0.27 | -1.25 – 0.70 | -0.55 |
| *Note. z* values > |1.96| are bolded.  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table S8 |  |  |  |
| *Final Model for the Probability to Initiate a Look-From from the Target Phrase* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 72 | 0.78 | 0.88 |
| Participant (Story Type) |  | 0.20 | 0.44 |
| Item (Intercept) | 44 | 0.07 | 0.27 |
| Item (Story Type) |  | > 0.001 | > 0.001 |
| Item (Group) |  | > 0.001 | > 0.001 |
| Item (Story Type ✕ Group) |   | 0.25 | 0.50 |
| **Fixed Effects** | *Estimate* | *95% CI* | *z* |
| Intercept | -2.23 | -2.52 – -1.94 | **-15.19** |
| Story Type (Literal vs. Ironic) | 0.22 | -0.13 – 0.58 | 1.23 |
| Group (Control vs. Training) | -0.04 | -0.56 – 0.47 | -0.17 |
| Time (Pre-Test vs Post-Test) | -0.31 | -0.59 – -0.04 | **-2.23** |
| Story Type ✕ Group  | -0.05 | -0.68 – 0.59 | -0.14 |
| Story Type ✕ Time | -0.54 | -1.09 – 0.01 | -1.91 |
| Group ✕ Time | 0.29 | -0.26 – 0.85 | 1.04 |
| Story Type ✕ Group ✕ Time | -0.12 | -1.22 – 0.98 | -0.22 |
| *Note. z* values > |1.96| are bolded.  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table S9 |  |  |  |
| *Final Model for First-Pass Reading Time on the Spillover Region* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 72 | 0.07 | 0.27 |
| Item (Intercept) | 44 | 0.13 | 0.36 |
| Item (Story Type) |  | 0.01 | 0.11 |
| Item (Group) |  | 0.003 | 0.06 |
| Item (Story Type ✕ Group) |  | 0.01 | 0.11 |
| Residual |   | 0.14 | 0.38 |
| **Fixed Effects** | *Estimate* | *95% CI* | *t* |
| Intercept | 6.66 | 6.54 – 6.79 | **103.77** |
| Story Type (Literal vs. Ironic) | 0.02 | -0.03 – 0.07 | 0.87 |
| Group (Control vs. Training) | 0.08 | -0.05 – 0.21 | 1.16 |
| Time (Pre-Test vs Post-Test) | -0.12 | -0.16 – -0.09 | **-7.01** |
| Story Type ✕ Group  | 0.07 | -0.01 – 0.15 | 1.73 |
| Story Type ✕ Time | -0.07 | -0.13 – 0.001 | -1.89 |
| Group ✕ Time | -0.01 | -0.08 – 0.06 | -0.27 |
| Story Type ✕ Group ✕ Time | -0.07 | -0.20 – 0.07 | -0.97 |
| *Note.* t values > |1.96| are bolded.  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table S10 |  |  |  |
| *Final Model for Probability to Look-Back to the Critical Context* |
| **Random Effects** | *n* | Variance | *SD* |
| Participant (Intercept) | 72 | 1.14 | 1.07 |
| Item (Intercept) | 44 | 0.23 | 0.48 |
| Item (Story Type) |  | 0.62 | 0.79 |
| Item (Group) |  | 0.01 | 0.10 |
| Item (Story Type ✕ Group) |   | 0.96 | 0.98 |
| **Fixed Effects** | *Estimate* | *95% CI* | *z* |
| Intercept | 0.16 | 0.12 – 0.23 | **-10.82** |
| Story Type (Literal vs. Ironic) | 0.87 | 0.61 – 1.24 | -0.77 |
| Group (Control vs. Training) | 0.93 | 0.52 – 1.66 | -0.24 |
| Time (Pre-Test vs Post-Test) | 0.62 | 0.48 – 0.80 | **-3.74** |
| Story Type ✕ Group  | 1.15 | 0.62 – 2.12 | 0.44 |
| Story Type ✕ Time | 0.63 | 0.38 – 1.04 | -1.82 |
| Group ✕ Time | 0.75 | 0.46 – 1.24 | -1.11 |
| Story Type ✕ Group ✕ Time | 1.65 | 0.61 – 4.46 | 0.98 |
| *Note. z* values > |1.96| are bolded.  |

**References**

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. <https://10.1016/j.jml.2007.12.005>

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. <https://doi.org/10.18637/jss.v067.i01>

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Metholdological), 26*, 211–252. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*, 389–411. <https://doi.org/10.1037/met0000159>

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Mächler, M., & Bolker, B.M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling*. The R Journal, 9*, 378–400. <https://doi.org/10.32614/RJ-2017-066>

Hartig, F. (2022). DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. Retrieved from: [https://cran.r-project.org/package=DHARM](https://cran.r-project.org/package%3DDHARM)

Hyönä, J., Lorch, R. F., & Rinck, M. (2003). Eye movement measures to study global text processing. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind’s eye: Cognitive and applied aspects of eye movement research* (pp. 313–334). Elsevier Science. <https://doi.org/10.1016/B978-044451020-4/50018-9>

Häikiö, T., Heikkilä, T. T., & Kaakinen, J. K. (2018). The effect of syllable-level hyphenation on reading comprehension: Evidence from eye movements. *Journal of Educational Psychology, 110*, 1149–1159. [https://doi.org/10.1037/edu0000261](https://psycnet.apa.org/doi/10.1037/edu0000261)

R Core Team (2021). R: A language and environment for statistical computing. Retrieved from: <http://www.R-project.org>

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110,* 104038. <https://doi.org/10.1016/j.jml.2019.104038>