**Appendix 1: Summary of Studies Included in Sakai and Moorman’s (2018) Meta-Analysis**

As shown in the summary table below, 9 out of 18 studies (50%) included in Sakai and Moorman’s (2018) meta-analysis were also included in the current meta-analysis. This means that 9 out of 31 studies (about 30%) came from Sakai and Moorman (2018). A notable improvement on the previous meta-analysis is related to the issue of “apples and oranges.” One ubiquitous problem with meta-analysis is that combining effect sizes from studies whose characteristics are substantially different makes it challenging to interpret the results. To respond to this issue, our focus was exclusively on HVPT studies which implemented a controlled perception training procedure (i.e., identification training) involving three important components of HVPT according to Thomson (2018, 2022): (1) multiple talkers, (2) multiple contexts, and (3) trial-by-trial feedback. Following this strict selection criteria, we ensured the consistency of the studies included in the current meta-analysis and needed to exclude 9 studies included in Sakai and Moorman (2018): three studies involved only a single talker; four studies adopted training procedures that substantially deviated from a commonly adopted perception procedure (e.g., listening to a text containing target sounds, listening to a recorded speech of other students & evaluating them for nativelikeness).

**Summary Table of 18 Studies Included in Sakai and Moorman (2018)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  No | Study | Type of publication | Included in the current meta-analysis | Why excluded? | Different talkers? | Description of training/treatment |
| 1 | Anderson (2011) | PhD thesis | No | No sufficient statistical information | Yes | Identification training |
| 2 | Bradlow et al. (1997) | Journal Article | Yes |   | Yes | Identification training |
| 3 | Counselman (2010) | PhD thesis | No | Not HVPT task | Yes  | Listening to & evaluating other students' productions |
| 4 | Gomez Lacabex & Garcia Lecumberri (2010) | Conference Proceedings | No | Single talker; Not HVPT task | No | Asked to induce rules after listening |
| 5 | Han (2002) | Journal Article | No | Not HVPT task | Yes | Instruction+Perception practice |
| 6 | Hazan et al. (2005) | Journal Article | Yes |   | Yes | Identification training |
| 7 | Herd et al. (2013) | Journal Article | Yes |   | Yes | Identification training |
| 8 | Huensch (2013) | PhD thesis | Yes (Huensch & Tremblay, 2015) |   | Yes | Identification training |
| 9 | Lambacher et al. (2005) | Journal Article | Yes |   | Yes | Identification training |
| 10 | Lengeris (2009) | PhD thesis | Yes (Lengeris & Hazan, 2010) |   | Yes | Identification training |
| 11 | Motohashi (2007) | PhD thesis | No | Single talker | No | Identification training with audiovisual or auditory input |
| 12 | Nobre-Oliveira (2007) | Conference Proceedings | No | No sufficient statistical information | Yes | Identification training  |
| 13 | Reis & Nobre-Oliveria (2007) | Conference Proceedings | No | Old-New stimulus status was not clear | Yes | Identification and Discrimination training |
| 14 | Soler-Urzua (2011) | Master thesis | No | Not HVPT task | Yes | Listening to stories read by a machine |
| 15 | Thomson (2007) | PhD thesis | Yes (Thomson, 2011) |   | Yes  | Identification training |
| 16 | Underbakke (1993) | Journal Article | No | Single talker | No | Identification and discrimination training with & without feedback |
| 17 | Wang (2002) | PhD thesis | Yes (Wang & Munro, 2004) |   | Yes | Identification training |
| 18 | Yeon (2004) | PhD thesis | Yes |   | Yes | Identification training |

**Appendix 2: Detailed Procedures and Equations for Calculating Within-Participant and Between-Participant Effect Sizes**

To compute the weighted mean effect size and conduct moderator analysis, we employed the Comprehensive Meta-Analysis (Version 3.3). We relied on Hedges’ g as the basic unit of analysis, the transformed version of Cohen’s d corrected for bias in small samples. To answer RQ1 regarding the overall effectiveness of HVPT for production improvement, we conducted two separate analyses according to different study designs: within-participant effect sizes (i.e., the mean difference between pretest and posttest scores) and between-participant effect sizes (i.e., the mean difference in gain scores between HVPT treatment and control groups).

Regarding the calculation of the effect size for the within-participant data, studies reporting sample means and standard deviations for pretest and posttest performance were extracted from a pool of descriptive data. The gain score was calculated by subtracting pretest from posttest score, and the standard deviation for the gain score was computed using **Equation 1**. Because the pretest-posttest correlation was needed to compute the standard deviation for the gain score, we imputed the pretest-posttest correlation for perception data (r = .60) calculated previously in Authors (XXXX) to estimate the standard deviation. For studies not reporting sample means and standard deviations for pretest and posttest performance, t values or sample mean differences and standard deviations for the differences were used to calculate the effect sizes.

Several studies reported multiple descriptive data from the same participants (the use of multiple elicitation tasks such as word reading and sentence reading, outcome measures such as identification and transcription, prompt modality, target phones, and test item types). Thus, such multiple scores except outcome measures were averaged to yield a single score to avoid violating the requirement of independence of observations. To take account of the nested structure of multiple data and to avoid losing information by averaging effect sizes, we averaged the observed effect sizes and aggregated the sampling variances for the effect sizes employing the procedure suggested by Borenstein et al. (2009) using **Equation 2**. Since this equation requires the correlation between the multiple effect sizes, we followed a previous meta-analysis of HVPT studies (Zhang et al., 2021) and imputed a correlation of .50 for aggregating dependent measures. As for multiple outcome measures, we followed Sakai and Moorman (2018) to select the elicited production with native speaker identifications as the representative task. Lastly, as suggested by Boreinstein et al. (2009), we used **Equation 3** to adjust the standard deviations for the gain scores using the pretest-posttest correlation (Borenstein et al., 2009). The resulting standard deviations within groups were used to compute the standardized mean difference for independent groups, later converted to Hedges’ g (see **Equation 4**). A total of 43 independent effect sizes were available for this analysis.

Regarding the calculation of the effect size for the between-participant data, we extracted studies reporting the sample means and standard deviations for both treatment and control groups. The group mean difference was calculated by subtracting the control-gain scores from the treatment-gain scores. The pooled standard deviation across the two groups was calculated using **Equation 5** based on the standard deviations for the gain scores from each group calculated using **Equation 1**. When two or more standard deviations were available from multiple independent treatment groups, the pooled standard deviation across groups was calculated using **Equation 5**. The group mean difference of gain scores was standardized using the pooled standard deviations for the two groups. The resulting standardized mean difference was converted to Hedges’ g (see **Equation 6**). As a result, 17 independent effect sizes were available for this analysis.

**Equation 1**

$$ S\_{diff}=\sqrt{SD\_{pre}^{2}+SD\_{post}^{2}-2×r×SD\_{pre}×SD\_{post}}$$

This equation was used to calculate the standard deviation of the gain scores (*Sdiff*) from the standard deviations of pretest and posttest scores and their correlation.

**Equation 2**

$$V\_{mean}=\left(\frac{1}{m}\right)^{2}\left(\sum\_{i=1}^{m}V\_{i}+\sum\_{i\ne j}^{}\left(r\_{ij}\sqrt{V\_{i}}\sqrt{V\_{j}}\right)\right)$$

This equation was used to calculate the variance of the mean of two or more correlated variables (*Vmean*) with *m* as the number of variances (*V*) (see Borenstein et al., 2009, p. 228).

**Equation 3**

$$S\_{within}=\frac{S\_{diff}}{2\sqrt{1-r}}$$

This equation was used to calculate the standard deviation for independent groups (*Swithin*) with the standard deviation for the gain scores (*Sdiff*) and the pretest-posttest correlation (*r* = 0.6).

**Equation 4**

$$Cohen^{'}s d\_{s}=\frac{M\_{diff}}{S\_{within}}$$

This equation was used to calculate the standardized mean difference between pretest and posttest scores adjusted for the pretest-posttest correlation (*ds*). Cohen’s *ds* was converted to Hedges’ *g* using a correction factor, called *J* with the degrees of freedom (*df* = *n* – 1):

$$J=1 -\frac{3}{4df-1}$$

$$Hedges^{'}g=d\_{s}×J$$

**Equation 5**

$$S\_{pooled}=\sqrt{\frac{\left(n\_{1}-1\right)SD\_{1}^{2}+\left(n\_{2}-1\right)SD\_{2}^{2}+\cdots +\left(n\_{k}-1\right)SD\_{k}^{2}}{n\_{1}+n\_{2}+\cdots +n\_{k}-k}}$$

This equation was used to calculate the pooled standard deviation (*Spooled*) for multiple independent groups.

**Equation 6**

$$Cohen^{'}s d\_{s}=\frac{M\_{diff}}{S\_{pooled}}$$

The group mean difference of gain scores (*Mdiff*) was standardized using *Spooled* to calculate Cohen’s *ds* and it was converted to Hedges’ *g* using a correction factor (*J*) with the degrees of freedom (*df* = *n1* + *n2* – 1):

$$J=1 -\frac{3}{4df-1}$$

$$Hedges^{'}g=d\_{s}×J$$

**References**

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187–224. https://doi.org/10.1017/S0142716417000418

Zhang, X., Cheng, B., & Zhang, Y. (2021). The role of talker variability in nonnative phonetic learning: A systematic review and meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802–4825. https://doi.org/10.1044/2021\_JSLHR-21-00181

**Appendix 3: Full References for 31 Studies in Meta-Analysis**

Aliaga-Garcia, C. (2017). *The effect of auditory and articulatory phonetic training on the perception and production of L2 vowels by Catalan-Spanish learners of English* [Unpublished doctoral dissertation]. Universitat de Barcelona.

Alves, U.K., & Luchini, P.L. (2017). Effects of perceptual training on the identification and production of word-initial voiceless stops by Argentinean learners of English. *Ilha Do Desterro*, 70(3), 15 – 32. https://doi.org/10.5007/2175-8026.2017v70n3p15

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

Carlet, A. (2017). *L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study* [Unpublished doctoral dissertation]. Universitat Autònoma de Barcelona.

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ*, *7*, e7191. https://doi.org/10.7717/peerj.7191

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495–522. <https://doi.org/10.1017/S0142716403000250>

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, 47(3), 360 – 378. <https://doi.org/10.1016/j.specom.2005.04.007>

Herd, W., Jongman, A., & Sereno, J. (2013). Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish. *The Journal of the Acoustical Society of America*, *133*(6), 4247–4255. <https://doi.org/10.1121/1.4802902>

Huensch, A (2013). *The perception and production of palatal codas by Korean L2 learners of English* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

 Hwang, H., & Lee, H.-Y. (2015). The effect of high variability phonetic training on the production of English vowels and consonants. In the Scottish Consortium for ICPhS (Ed.), *Proceedings of the 18th international Congress of Phonetic Sciences*. Paper number 486. Glasgow, UK: Glasgow University.

Iino, A., & Thomson, R.I. (2018). Effects of web-based HVPT on EFL learners’ recognition and production of L2 sounds. In P. Taalas, J. Jalkanen, L. Bradley & S. Thouesny (Eds.), *Future-proof CALL: Language learning as exploration and encounters – short papers from EUROCALL 2018* (pp. 106 – 111). <https://research-publishing.net/book?10.14705/rpnet.2018.26.9782490057221>

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145–160. <https://doi.org/10.1017/S0142716411000300>

Kangatharan, J., Giannakopoulou, A., & Uther, M. (2021) The effect of HVP training in vowel perception on bilingual speech production. *Research in English Language Teaching*, 2(1), 15 – 32. doi: 10.18488/journal.179.2021.21.15.32

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247. https://doi.org/10.1017/S0142716405050150

Lee, A., & Lyster, R. (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics*, 38, 371 – 393. https://doi.org/10.1017/S0142716416000254

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *The Journal of the Acoustical Society of America*, 128, 3757 – 3768. <https://doi.org/10.1121/1.3506351>

Li, Y. (2015). *Audio-visual training effect on L2 perception and production of English /θ/-/s/ and /ð/-/z/ by Mandarin speakers* [Unpublished doctoral dissertation]. Newcastle University.

Lopez-Soto, T., & Kewley-Port, D., (2009). Relation of perception training to production of codas in English as a second language. *Proceedings of Meetings on Acoustics*, 6, 1 – 15. <https://doi.org/10.1121/1.3262006>

Macdonald, R.M.M. (2012). *Counteracting age related effects in L2 acquisition: Training to distinguish between French vowels* (Unpublished doctoral dissertation). The University of Edinburgh.

Mokari, P.G., & Werner, S. (2017). Perceptual training of second-language vowels: Does musical ability play a role? *Journal of Psycholinguistic Research*, 47, 95 – 112. <https://doi.org/10.1007/s10936-017-9517-8>

Okuno, T., & Hardison, D.M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology*, 20(2), 61 – 80. Retrieved from <http://llt.msu.edu/issues/june2016/okunohardison.pdf>

Reyes, Y.P., & Hazan, V., (2021). English vowel perception by non-native speakers: Impact of audio and visual training modalities. Onomazein: Journal of Linguistics, Philology and Translation, 51, 111 – 136. <https://doi.org/10.7764/onomazein.51.04>

Thomson, R.I. (2007). *Modeling L1/L2 interactions in the perception and production of English vowels by Mandarin L1 speakers: A training study* (Unpublished doctoral dissertation). University of Alberta.

Thomson, R.I., & Derwing, T. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le, E. Simpson, & S. Vo (Eds.) *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference* (pp. 88 – 97). Ames, IA: Iowa State University.

Trakantalerngsak, T. (2016). *The effect of perceptual training on the learning of Japanese fricative and affricate contrasts by native Thai learners of Japanese* (Unpublished doctoral dissertation). Osaka University.

Ueda, R., & Hashimoto, K. (2019). Perceptual training in a classroom setting: Phonemic category formation by Japanese EFL learners. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 237–249). Iowa State University.

Wang, X. (2002). *Training Mandarin and Cantonese speakers to identify English vowel contrasts: Long-term retention and effects on production* (Unpublished doctoral dissertation). Simon Fraser University.

Wang, Y., Jongman, A., & Serano, A. (2003). Acoustic and perceptual evaluation of Mandarin tone production before and after perceptual training. *The Journal of the Acoustical Society of America*, 113, 1033 – 1043. https://doi.org/10.1121/1.1531176

Wong, W.S. (2013). *Training the perception and production of English vowels /I/-/i:/, /e/-/æ/ and /ʊ/ -/u:/by Cantonese ESL Learners in Hong Kong* (Unpublished doctoral dissertation). Chinese University of Hong Kong.

Yang, R., Nanjo, H., & Dantsuji, M. (2021). Self adaptive phonetic training for Mandarin nasal codas. *Computer-Assisted Language Learning Electronic Journal*, *22*(1), 391–413.

Yeon, S.-H. (2004). *Teaching English word-final alveolopalatals to native speakers of Korean* [Unpublished doctoral dissertation]. University of Florida.