Puanlayıcı Tecrübesi ve Kompozisyon Kalitesinin Puanlayıcı Davranışı ve Kompozisyon Puanları Üzerindeki Etkisi

Özet

Bu proje, proje yürütücüsü danışmanlığında proje araştırmacısı tarafından hazırlanmakta olan doktora tezi araştırmasıdır. Bu çalışma, temel olarak puanlayıcı tecrübesinin ve kompozisyon kalitesinin puanlayıcı davranışı ve kompozisyon puanları üzerindeki etkilerini araştırmayı hedeflemektedir. Öncelikle, tecrübeli ve tecrübesiz puanlayıcılar üzerinde yapılacak araştırmayla mesleki deneyimin kompozisyon puanlarının değişkenliği ve güvenirliği üzerindeki etkisi saptanacaktır. Ayrıca, deneyimli ve deneyimsiz puanlayıcıların farklı kalitedeki kompozisyonları puanlarken sergiledikleri puanlayıcı davranışları ve karar verme stratejileri gözlemlenecektir. Bununla birlikte, yukarıda bahsedilen faktörlerin, yabancı dil olarak İngilizce öğrenen öğrencilerin kompozisyon puanlarının değişkenliğini ve güvenirliğini ne derecede etkilediği ölçülecektir.

Yazma performansını değerlendirme sürecinin güvenirlik, geçerlilik ve tarafsızlık açısından tartışmalı bir konu olmasından ötürü, puanlayıcının yazma becerisinin öğretimi ve değerlendirilmesi alanındaki mesleki tecrübesi, kompozisyon kalitesi, değerlendirme aşamasında sergilenen puanlayıcı davranışları ve öğrencilerin kompozisyonlarına verilen puanlar arasındaki çoklu etkileşimleri incelemek büyük önem arz etmektedir. Tecrübeli ve tecrübesiz puanlayıcılar düşük ve yüksek kalitedeki kompozisyonlar ile ilgili farklı algılara sahip olabilmektedirler. Bu nedenle, puanlayıcıların mesleki deneyimlerine bağlı olarak sergiledikleri puanlayıcı davranışları ve karar verme süreçleri kompozisyon puanlarının değişkenliğine yol açabilecek farklılaşmalarla sonuçlanabilir. Bu yüzden, bu çalışma mesleki tecrübeyi göz önüne alarak, puanlayıcıların bilişsel yapısının ve kompozisyon kalitesinin puan değişkenliği ve yazma performansı değerlendirme güvenirliği üzerindeki etkilerini derinlemesine inceleyecektir. Ayrıca, bu proje yukarıda bahsedilen faktörler arasındaki çok yönlü etkileşimi ve bu faktörlerin adil olmayan değerlendirmeyi ne ölçüde etkilediğini inceleyecektir.

Araştırmacı, farklı yükseköğretim kurumlarında görev yapan ve yabancı dil olarak İngilizce öğreten toplam 32 öğretim elemanına ulaşmıştır. Bu çalışmaya katılacak olan öğretim elemanlarının 15'i Bursa Teknik Üniversitesi'nde (BTÜ) görev yapan öğretim elemanlarından seçilmiş olup, kalan 17 katılımcı ise Türkiye'nin çeşitli üniversitelerinde çalışmaktadır. Bu dağılım, sonuçların kurumsal ve ulusal olmak üzere iki farklı perspektiften yorumlanmasını sağlayacaktır. Araştırmacı, öğretim elamanlarına düşük ve yüksek kalitede olmak üzere 50 adet kompozisyon verecektir. Söz konusu kompozisyonlar, Çanakkale Onsekiz Mart Üniversitesi Eğitim Fakültesi Yabancı Diller Eğitimi Anabilim Dalı İngilizce Öğretmenliği Anabilim Dalı'nda proje yürütücüsü tarafından verilmekte olan İleri Okuma ve Yazma Becerileri Dersi'ne kayıtlı öğrencilerden toplanmıştır.

Bu çalışmada karma araştırma yaklaşımı benimsenmiş olduğundan, puanlayıcılardan hem nitel hem de nicel veri toplanacaktır. Öncelikle, her bir puanlayıcı tarafından 50 adet kompozisyona verilecek olan ve toplamda elde edilecek 1600 adet kompozisyon puanı çalışmanın sayısal verisini oluşturacaktır. Puanlayıcılar, kompozisyonları değerlendirirken bu proje kapsamında adaptasyon çalışmaları yapılacak olan analitik bir ölçek kullanacaklardır. Bu veriler, genellenebilirlik kuramı (G Kuramı) çerçevesinde istatistiksel analize tâbi tutulacaktır. İkinci olarak, puanlama işlemini tamamladıktan

hemen sonra puanlayıcılardan özgeçmişleriyle ilgili bilgi vermeleri istenecektir. Bu veriler ise, puanlayıcıların mesleki tecrübeleri, cinsiyetleri ve eğitim geçmişlerine göre sınıflandırılmasında kullanılacak ve nitel ve nicel verilerin yorumlanmasında yardımcı olacaktır. Puanlayıcıların bilişsel dünyasını ve karar verme süreçlerini araştırmak adına, sesli düşünme protokolleri (think-aloud protocols) ile nitel veri toplanacaktır. Her bir puanlayıcı, kompozisyonların 16 tanesini sesli düşünme yöntemi ile puanlayacaktır ve tolamda 512 ses kaydı elde edilecektir. Puanlayıcılar kompozisyonları değerlendirirken düşüncelerini nasıl dile getirecekleri ile ilgili eğitim alacaklardır. Puanlayıcıların hangi karar verme süreçlerinden geçtiklerini ve kompozisyonları hangi açılardan değerlendirdiklerini belirlemek için, kompozisyonları değerlendirirken dile getirecekleri düşünceleri kaydedilecek, bu kayıtlar metne dökülecek ve daha sonra analiz edilecektir. Sesli düşünme protokollerinden elde edilecek olan verileri düzenlemek için bir kodlama şeması kullanılacaktır. Kodlama işleminin güvenirliğini kontrol etmek adına, yazma performansı değerlendirme alanında uzman olan bağımsız bir araştırmacı, sesli düşünme protokollerinden elde edilen verinin %15'lik bir bölümünü kodlayacaktır.

Puanlayıcı deneyimi ile birlikte pek çok faktörün puan güvenirliği ve değişkenliği üzerindeki etkisini araştıran literatürdeki çalışmaların çoğu, İngilizcenin ikinci dil statüsüne sahip olduğu ortamlarda yapılmıştır. Bu proje, puanlayıcıların kompozisyon değerlendirme deneyimlerinin puanlayıcı davranışları ve puan değişkenliği üzerindeki etkilerini, İngilizcenin yabancı dil statüsüne sahip olduğu bir ortamda araştıracağından önemlidir. Ayrıca, kompozisyon kalitesi ve kompozisyon puanları arasındaki ilişki de literatürde yeterince incelenmemiş olup, konuyla ilgili yapılan çalışmaların çoğu İngilizcenin ikinci dil statüsünde olduğu ortamlarla ilgili bilgi vermektedir. Bu çalışma, kompozisyon kalitesinin değerlendirme puanları üzerindeki etkisini araştırarak literatürdeki araştırma boşluğunu doldurmayı hedeflemektedir. Nicel çalışmaların birçoğu klasik test teorisiyle temellendirilmiştir; ancak bu araştırma, puanlayıcı değişkenliği ve kompozisyon değerlendirme güvenirliğini saptamak için daha karmaşık bir teori olan G kuramını benimsemiştir. Ayrıca, tecrübeli ve tecrübesiz puanlayıcıların değerlendirme kriterleri ve puanlama süreçlerini karşılaştırmak maksadıyla sesli düşünme protokollerinin kullanılacak olması bu projeyi önemli kılmaktadır. Bu çalışmada elde edilecek sonuçların, yukarıda bahsedilen faktörlerden kaynaklanan yazma performansı puanlama problemlerine ve adil olmayan değerlendirmelere ışık tutması beklenmektedir. Bu bağlamda, projenin araştırma ekibi Türkiye'de yer alan yükseköğretim kurumlarında İngilizce kompozisyonlara verilen puanlarının güvenirliğinden emin olmak ve ilerleyen zamanlarda Öğrenci Seçme ve Değerlendirme Merkezi (ÖSYM) tarafından yapılması planlanan İngilizce yazma sınavlarının değerlendirme güvenirliğini artırmak adına puanlayıcı tecrübesi ve kompozisyon kalitesinden kaynaklanan sorunlara çözüm önerileri sunmayı hedeflemektedir. Ayrıca, araştırma bulguları neticesinde ideal puanlayıcı profili resmedilecek ve bir puanlayıcı eğitimi modeli sunulacaktır.

Anahtar Kelimeler: G Kuramı, kompozisyon kalitesi, puan değişkenliği, puanlayıcı davranışı, puanlayıcı tecrübesi, sesli düşünme protokolü, yazma performansı değerlendirme

Amaç ve Hedefler

Amaçlar

Bu çalışmanın temel amacı, hem puanlayıcı tecrübesinin hem de kompozisyon kalitesinin puanlayıcı davranışı ve kompozisyon puanları üzenindeki etkisini araştırmaktır. Öncelikle, proje kapsamında tecrübeli ve tecrübesiz puanlayıcıların incelenmesiyle mesleki deneyimin, değerlendirme puanlarının güvenirliği ve değişkenliği açısından önemli olup olmadığı araştırılacaktır. İkinci olarak, bu çalışma tecrübeli ve tecrübesiz puanlayıcıların farklı kalitedeki kompozisyonları değerlendirirken sergiledikleri puanlayıcı davranışlarını ve karar verme stratejilerini gözlemlemeyi amaçlamaktadır. Son olarak, yukarıda bahsedilen değişkenlerin, yabancı dil olarak İngilizce öğrenen öğrencilerin kompozisyonlarına verilen değerlendirme puanlarının güvenirliğine ne ölçüde etki ettiği hesaplanacaktır.

Hedefler

Yukarıda belirtilen amaçlar doğrultusunda, bu çalışma yabancı dil olarak İngilizce öğrenen öğrencilerin, yükseköğretim kurumlarında yazma becerilerinin güvenilir bir şekilde değerlendirilmesi için çözümler sunmayı hedeflemektedir. Bu anlamda, çalışmanın bulguları İngilizce hazırlık programlarında, ERASMUS gibi değişim programlarında ve İngiliz dili ile ilgili olan lisans bölümlerinde müfredatta yer alan İleri Okuma ve Yazma Becerileri ve Akademik Yazma gibi derslerde öğrencilerin yazılı performanslarının adil bir değerlendirmeye tabi tutulmasına yardımcı olacaktır. Böylelikle bu çalışma, hem kurumsal hem de ulusal düzeyde yükseköğretim kurumlarında, yazma performansı değerlendirme sürecinin standartlaştırılması ve güvenilir hale getirilmesi açısından çözümler üretmeyi hedeflemektedir. Ayrıca bu çalışma, ÖSYM tarafından yapılmakta olan Yabancı Dil Sınavına (YDS) ilerleyen zamanlarda entegre edilmesi planlanan yazma performansı sınavlarının değerlendirilmesinde ihtiyaç duyulacak olan puanlayıcı kriterlerinin belirlenmesine de katkı sağlamayı hedeflemektedir. Buna ek olarak, bu projede elde edilecek bulgular neticesinde, bir puanlayıcı eğitimi modelinin tasarlanması ve başka bir proje ile ilgili öğretim elemanlarına puanlayıcı eğitimi verilmesi hedeflenmektedir.

Konu Kapsamı ve Sınırları

Yabancı dil olarak İngilizce öğrenen öğrencilerin kompozisyonlarını değerlendirirken farklı değerlendirme protokolleri uygulanmakta ve ortak bir yol izlenememektedir. Bazı üniversitelerde öğrenci ödevleri, hiçbir değerlendirme ölçeği kullanılmadan tek bir puanlayıcı tarafından notlanırken; bazı üniversitelerdeyse aynı kompozisyonlar iki farklı puanlayıcı tarafından değerlendirilmektedir. Bazı kurumlar öğrenci performansına adil bir puan vermek için, anonim değerlendirme uygulamasını benimserken; açık değerlendirme sürecinin yürütüldüğü kurumlardaysa değerlendirme sürecinde öğrenci kimliklerinin etkisi altında kalınabilmektedir. Aynı kurum içerisinde ve/veya farklı kurumlarda uygulanan ve benzerlik göstermeyen puanlama yöntemleri, adil olmayan değerlendirmeleri beraberinde getirmektedir. Bu yüzden, üniversitelerde öğrencilerin İngilizce yazma performansını değerlendirirken standart ve güvenilir yöntemlerin kullanılmasına ihtiyaç duyulmaktadır.

Yazma performansı değerlendirmesi, güvenirlik, geçerlilik ve tarafsızlık açısından sorunları beraberinde getiren bir süreçtir. Bu yüzden, puanlayıcının mesleki tecrübesi,

kompozisyon kalitesi, puanlayıcı davranışları, puanlayıcıların karar verme süreçleri gibi faktörlerin ele alınması ve bu faktörler arasındaki etkileşimin irdelenmesi, kompozisyon değerlendirme puanlarının güvenirliği açısından önem arz etmektedir. Tecrübeli ve tecrübesiz puanlayıcılar düşük ve yüksek kalitedeki öğrenci kompozisyonlarını farklı şekilde algılayabilirler. Benzer şekilde, puanlayıcıların kompozisyon değerlendirme sürecinde sergiledikleri davranışlar ve takip ettikleri karar verme aşamaları da, yazma becerisi öğretimi ve kompozisyon değerlendirme açısından sahip oldukları mesleki tecrübeye bağlı olarak değişiklikler gösterebilir ve bu farklılaşma, kompozisyon puanlarının değişkenliği ile sonuçlanabilir. Bu yüzden bu araştırma, mesleki deneyim ve kompozisyon kalitesinin etkileşimi ve karar verme süreçleriyle birlikte puanlayıcının bilişsel yapısını ele alacaktır. Bu faktörler arasındaki ilişkinin ve etkileşimin de kompozisyon puanlarının değişkenliğine ne ölçüde yansıdığı araştırılacaktır. Ülkemizde G kuramı kullanılarak yazma performansı değerlendirmesi ile ilgili istatistiksel çalışmaların çok fazla yapılmamış olması ve kompozisyon kalitesinin değerlendirme puanlarının güvenirliğine ve değişkenliğine etki eden bir faktör olarak literatürde yeterince araştırılmamış olması bu çalışmayı değerli kılmaktadır.

Araştırma Soruları

Yukarıda belirtilen ana amaç doğrultusunda, bu çalışmada toplanacak olan nicel veri ışığında 4 (1 – 4), nitel veri ışığındaysa 2 (5 – 6) olmak üzere aşağıdaki toplam 6 araştırma sorusuna yanıt aranacaktır:

1. Düşük ve yüksek kalitedeki öğrenci kompozisyonlarına verilen analitik puanlar arasında anlamlı fark var mıdır?

2. Düşük ve yüksek kalitedeki öğrenci kompozisyonlarına tecrübeli ve tecrübesiz puanlayıcılar tarafından verilen analitik puanlar arasında anlamlı fark var mıdır?

3. Öğrenci kompozisyonlarına verilen analitik puanların farklılaşmasına en çok katkı sağlayan (göreceli olarak) değerlendirme puan değişkenliği kaynakları nelerdir?

4. Tecrübeli puanlayıcıların öğrenci kompozisyonlarına verdikleri analitik puanların güvenirliği (örneğin, kriter referanslı puan yorumlarının güvenirlik katsayıları) ile tecrübesiz puanlayıcıların verdiği puanların güvenirliği arasında anlamlı fark var mıdır?

5. Puanlayıcılar yabancı dil olarak İngilizce öğrenen öğrencilerin kompozisyonlarını analitik bir şekilde değerlendirirken hangi karar verme süreçlerinden geçmektedirler?

6. Mesleki tecrübe, İngilizce kompozisyonların değerlendirilmesinde puanlayıcıların karar verme süreçlerinde ve kompozisyonlarda dikkat ettikleri noktalar üzerinde bir etkiye sahip midir?

Literatür Özeti

İkinci dil olarak İngilizce yazma becerisini güvenilir bir şekilde değerlendirmek sorunlu bir süreçtir; çünkü pek çok değişken değerlendirme puanlarının güvenirliğine, geçerliliğine ve tarafsızlığına etki etmektedir (Breland, 1983; Han, 2013; Huang, 2007, 2008, 2009, 2011; Huang ve Foote, 2010). Güvenirlik sorununa yol açan puan değişkenliği; öğrenciler, puanlayıcı ve kompozisyon olmak üzere üç ana unsurdan kaynaklanmaktadır (Elorbany ve Huang, 2012; McColly, 1970). Örneklendirmek gerekirse; puanlayıcının ana dili, mesleki deneyimleri ve geçmişleri veya kompozisyon konusu, öğrencilerin İngilizce kompozisyon puanlarını etkileyebilir. Buna ek olarak, farklı puanlayıcılar tarafından aynı kompozisyona verilen değerlendirme puanları da puanlama biçimi, ölçek türü ve puanlayıcı eğitimine bağlı olarak değişiklikler gösterebilir (Barkaoui, 2008; Brown, 1991; Cumming, 1990; Huang, 2007, 2008, 2009, 2011; Huang ve Han, 2013; Lumley, 2005; Weigle, 1994, 2002).

Yukarıda bahsedilen faktörler arasında puanlayıcı değişkeni, yazma performansı değerlendirme sürecinin merkezinde yer almaktadır (Bachman, 1990; Huang, 2008, 2011; Huang ve Foote, 2010; Huot, 1990; Lim, 2011; Stalnaker ve Stalnaker, 1934). Puanlayıcılar; mesleki deneyimleri, ana dilleri, eğitim geçmişleri, beklentileri ve inançları ve öğrenci hatalarına karşı sahip oldukları hoşgörü seviyelerine kadar çok sayıda farklılıklar göstermektedirler (Huang, 2009; Weigle, 2002). Bu farklılıklar, birden fazla puanlayıcı tarafından aynı kompozisyona verilen puanların (inter-rater reliability) veya farklı zamanlarda aynı kompozisyona tek bir puanlayıcı tarafından verilen puanların (intra-rater reliability) değişkenlik göstermesine neden olmaktadır (Bachman, 1990; Homburg, 1984; Huang, 2008, 2009, 2011; Huot, 1990).

Puanlayıcı özellikleri arasında mesleki deneyimin değerlendirme sürecinde önemli bir rol oynadığı düşünülmektedir (Barkaoui, 2010b). Mesleki deneyimin kompozisyon puanları üzerindeki etkisi ile ilgili pek çok araştırma yapılmıştır ve bu araştırmalar çelişkili sonuçlar ortaya koymuştur (örneğin, Barkaoui, 2008; Cumming, 1990; Hamp-Lyons, 1996; Reid ve O'Brien, 1981; Shohamy, Gordon ve Kraemer, 1992; Song ve Caruso, 1996; Sweedler-Brown, 1985; Weigle, 1999). Reid ve O' Brien (1981), İngilizceyi ikinci dil olarak öğrenen öğrencilerin kompozisyonlarına 10 adet puanlayıcı tarafından verilmiş holistik puanların güvenirliğini ve geçerliliğini incelemişlerdir. Çalışmanın bulguları, mesleki deneyim ile puanlayıclar-arası (inter-rater) ve tek puanlayıcı güvenirliği (intra-rater) arasında pozitif korelasyon göstermiştir. Song ve Caruso (1996), uzman puanlayıcıların, öğrenci kompozisyonlarına verilen holistik puanlarda tecrübesiz puanlayıcılara göre daha hoşgörülü olduğu sonucuna varmışlardır. Nitekim her iki puanlayıcı grubu, analitik değerlendirme ile kompozisyonlara benzer puanları vermişlerdir. Cumming (1990), puanlayıcıların öğrenci kompozisyonlarını holistik yöntemle değerlendirirken öğrencilerin İngilizce dil yeterlilikleri ve yazma becerilerini birbirinden farklı bir şekilde ele alıp almadıklarını ve tecrübeli ve tecrübesiz puanlayıcıların değerlendirme sürecindeki davranış biçimlerini araştırmıştır. Araştırma sonuçlarına göre, her iki puanlayıcı grubu sergiledikleri karar verme davranışlarında anlamlı bir şekilde birbirlerinden farklılaşmışlardır. Buna ek olarak, tecrübeli ve tecrübesiz puanlayıcıların "içerik" ve "retorik organizasyon" açısından kompozisyonlara verdikleri puanlar anlamlı farklar sergilerken, her iki puanlayıcı grubunun "dil kullanımı" açısından yaptıkları değerlendirmeler benzerlik göstermektedir. Weigle (1999) puanlayıcı-kompozisyon türü (rater-prompt) arasındaki etkileşimi, tecrübeli ve tecrübesiz puanlayıcıların ikinci dil olarak İngilizce öğrenenlerin kompozisyonlarına verdikleri puanlar açısından ele almıştır. Çalışmanın bulgularına göre, tecrübesiz puanlayıcılar

grafik yorumlama türündeki kompozisyonlara tecrübeli puanlayıcılara göre daha az puan vermiştir. Nitekim bu fark puanlayıcı eğitimi sonrasında giderilmiştir. Bu çalışmaları takiben Barkaoui'nin (2008) çalışmasında, İngilizceyi ikinci dil olarak öğrenen öğrencilerin yazma performanslarına verdikleri analitik ve holistik puanlar açısından, 31 adet tecrübeli ve 29 adet tecrübesiz puanlayıcıyı incelenmiştir. Araştırma sonuçlarına göre, hem deneyimli hem de deneyimsiz puanlayıcılar, kompozisyonların sahip olduğu iletişim kalitesini (communicative quality) en önemli unsur olarak düşünmektedirler. Nitekim, nispeten daha az tecrübeye sahip olan puanlayıcıların, değerlendirme aşamasında daha hoşgörülü oldukları ve kompozisyonlardaki tartışma bölümlerine daha fazla önem verdikleri belirlenmiştir. Öte yandan, tecrübeli puanlayıcıların tecrübesizlere göre daha katı oldukları ve kompozisyonlardaki dil bilgisi kurallarına daha fazla önem verdikleri sonucuna varılmıştır. Bununla birlikte, tecrübeli puanlayıcıların puanlama aşamasında değerlendirme kriterlerine daha sadık kaldıkları da elde edilen sonuçlardandır. Shohamy ve diğerleri (1992) ise tecrübeli ve tecrübesiz puanlayıcıların, yabancı dil olarak İngilizce öğrenenlerin kompozisyonlarına vermiş oldukları puanları incelemiş ve yukarıda bahsedilen çalışmaların aksine puanlayıcılar arasında anlamlı bir fark bulamamıştır.

Bu araştırmada ele alınacak diğer bir etken olan kompozisyon kalitesinin, yazma performansı değerlendirme puanlarının değişkenliği ve güvenirliği üzerindeki etkisini araştıran çalışmaların sayısı oldukça kısıtlıdır. Var olan çalışmaların odak noktasında da öğrencilerin hedef dildeki yetkinliği ve anadilleri yer almaktadır (Baba, 2009; Brown, 1991; Huang, 2008; Huang, Han, Tavano ve Hairston, 2014). Brown (1991), analidi İngilizce olan öğrencilerin ve İngilizceyi ikinci dil olarak öğrenen öğrencilerin yazmış oldukları kompozisyonlara verilen puanları incelemiştir. Bu çalışmanın sonuçlarına göre, her iki grup öğrencilerin yazılı ürünlerine verilen puanlar arasında anlamlı bir fark bulunmamıştır. Bununla birlikte Huang (2008), İngilizceyi ikinci dil olarak öğrenen öğrencilerin hedef dil bilgisindeki eksikliklerinden dolayı analdili İngilizce olan öğrencilere göre kompozisyonlarına daha az puan aldıkları sonucuna varmıştır. Benzer şekilde, Baba (2009) kelimelerin uygun bir şekilde kullanımının (appropriate use of lexical items), yazma performansına ve değerlendirme puanlarına olumlu açıdan katkı sağladığını öne sürmüştür. Son dönemlerde yapılan bir çalışmada, Huang ve diğerleri (2014) ikinci dil olarak İngilizce öğrenen öğrencilerin yazdıkları kompozisyonların kalitesinin, değerlendirme puanlarının değişkenliği ve güvenirliği üzerindeki etkisini incelemiştir. Çalışmanın bulgularına göre, düşük ve yüksek kalitedeki öğrenci kompozisyonlarına verilen holistik puanların standart sapması büyük çıkmış; ancak orta derecede kaliteli olan öğrenci kâğıtlarına verilen puanların standart sapması daha küçük bulunmuştur. Analitik değerlendirme açısından, düşük, orta ve yüksek kalitedeki kompozisyonlar için daha küçük standart sapmalar elde edilirken, bu kâğıtlar için daha yüksek ortalama puanlar gözlemlenmiştir.

Daha öncede belirtildiği gibi, diğer faktörlerle birlikte puanlayıcı tecrübesinin kompozisyon puanlarının değişkenliği ve güvenirliği üzerindeki etkisini araştıran çalışmaların çoğu İngilizcenin ikinci dil olarak öğretildiği ortamlarda yürütülmüştür (örneğin, Barkaoui, 2008, 2010b; Cumming, 1990; Hamp-Lyons, 1996; Reid ve O'Brien, 1981; Song ve Caruso, 1996; Weigle, 1999). Bu yüzden, bu çalışma, puanlayıcı deneyiminin puanlama güvenirliği üzerindeki etkisini, İngilizcenin yabancı dil olarak öğretildiği bir ortamda, yani Türkiye'de araştıracak olmasıyla önemlidir. Ayrıca, kompozisyon kalitesinin, diğer bir ifadeyle öğrencinin hedef dildeki yetkinliğinin, değerlendirme puanları üzerindeki etkisi araştırılarak, az sayıda ve çoğunun İngilizcenin

ikinci dil statüsüne sahip olduğu durumlarla sınırlı olan çalışmalara Türkiye örnekleminde yürütülecek bu çalışma ile katkı sağlanacaktır.

Özgün Değer

İyi bilindiği üzere, insan eliyle yapılan değerlendirmeler pek çok faktörün etkisi altında kalmaktadır ve bu faktörler puanlayıcı özellikleri olarak bilinmektedir (Eckes, 2012). Yapılan deneysel çalışmalar (Barkaoui, 2008; Cumming, 1990; Hamp-Lyons, 1996; Reid ve O'Brien, 1981; Rinnert ve Kobayashi, 2001; Shohamy, Gordon ve Kraemer, 1992; Song ve Caruso, 1996; Weigle, 1999), puanlayıcı deneyiminin değerlendirme puanları üzerinde önemli bir rol oynadığını bildirse de, elde edilen sonuçlar tecrübenin, puanlayıcılar-arası (inter-rater) ve tek puanlayıcı güvenirliği (intra-rater) (Cumming, 1990; Reid ve O'Brien, 1981; Shohamy ve diğ., 1992), puanlayıcı katılığı (rater severity) (Song ve Caruso, 1996), veya puanlama yöntemleri (scoring methods) (Barkaoui, 2008, 2010a) ile pozitif bir korelasyona sahip olduğu yönünde bulgular ortaya koymaktadır. Bu açıdan, araştırmanın puanlayıcı tecrübesinin değerlendirme puanları ve punlayıcı davranışları üzerindeki etkilerini ele alacak olması ve Türkiye'de yükseköğretim kurumlarında var olan İngilizce yazma performansı ile ilgili ölçme ve değerlendirme sorunlarına ışık tutmayı hedeflemesi, bu projeyi değerli kılan faktörlerin başında sıralanabilir. Bu araştırmanın farklı tecrübelere sahip olan puanlayıcıların değişen kalitedeki öğrenci kompozisyonlarıyla ilgili algılarını inceleyecek ve kompozisyonlara verdikleri reaksiyonları analiz edecek olması, projeye özgün değer katan diğer bir unsurdur. Bu şekilde, tecrübeye bağlı olarak öğrenci kompozisyonlarına tutarlı puanlar veren değerlendirmecinin profili ve bilişsel dünyası resmedilecektir. Böylelikle, ÖSYM tarafından yapılması planlanan ve öğrenciler için yüksek derecede önem arz eden yabancı dil olarak İngilizce yazma sınavları için değerlendirmeci kriterleri açısından öneriler sunulacaktır.

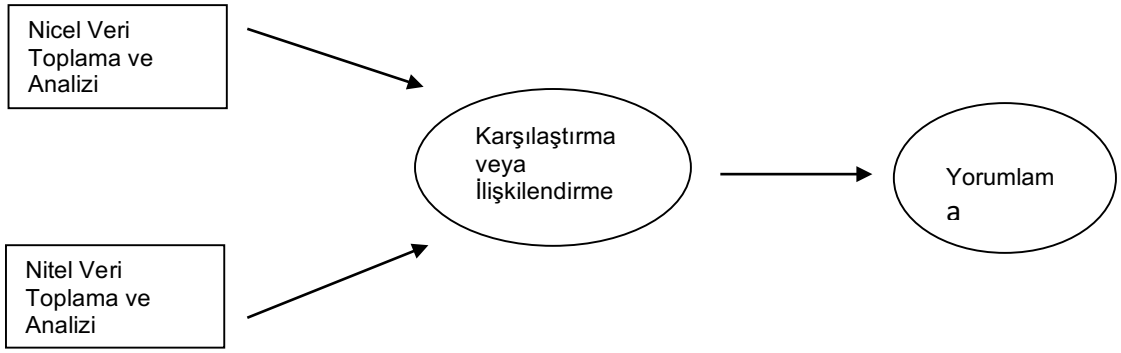
Ayrıca, araştırmanın teorik çerçevesinin de bu projeyi değerli kıldığı söylenebilir. Nicel araştırmaların çoğu klasik test teorisine dayandırılarak yapılmıştır; ancak yapılan analiz çerçevesinde sadece tek bir değişken kaynağına açıklama getirdiğinden bu teorinin zayıf bir kuram olduğu düşünülmektedir (Huang, 2008, 2012; Linn ve Burton, 1994). Bu yüzden, bu projede nicel verilerin analizi için G Kuramı kullanılacaktır. G Kuramı, bu proje kapsamında kompozisyon puanlarının sergilediği değişkenliğe neden olan hata kaynaklarına açıklık getirmede ve her bir hata kaynağının değerlendirme puanlarına ne ölçüde katkı sağladığını belirlemede kullanılan karmaşık bir test modelidir.

Bu projeye özgün değer katan diğer bir unsur ise, proje sonunda Türkiye'de yükseköğretim kurumlarında tarafsızlık ve güvenirlik açısından var olan kompozisyon değerlendirme problemlerine ışık tutulacak olmasıdır. Puanlayıcıların tecrübeleriyle ilişkili olarak, bilişsel yapıları incelenecek ve karar verme stratejileri tespit edilecektir. Elde edilen bu bulgular sayesinde ideal puanlayıcı profili resmedilecek ve yeni bir puanlayıcı eğitimi modeli sunulacaktır. Bu şekilde kompozisyon değerlendirme sürecinin standartlaşması adına önemli bir adım atma fırsatı elde edilecektir.

Yöntem

Karma bir yönteme (mixed method) sahip olan bu keşfe dayalı deneysel çalışmada, farklı türden araçlarla veri toplanacaktır. Karma araştırma yöntemi, nitel ve nicel verilerle araştırma bulgularını desteklemekte ve daha geçerli sonuçlar ortaya çıkarmaktadır (Dörnyei, 2007). Bu bağlamda, kompozisyon puanlarına ek olarak, sesli düşünme

protokolleri ile puanlayıcılardan nitel veri de toplanacaktır. Karma araştırma metodu kapsamında, projede yakınsayan paralel karma yöntem deseni (convergent paralel design, Creswell, 2011, Bakınız Şekil 1) kullanılacaktır. Araştırma desenine göre, nitel ve nicel veriler eş zamanlı toplanacak; ancak ayrı ayrı analiz edilecektir. Bu yöntemdeki temel varsayım nitel ve nicel verilerin farklı türde bilgiler sağlayacak olmasıdır. Araştırma sorularına ışık tutmak adına nitel ve nicel veriler eşit derecede önem arz etmektedir.



Şekil 1. Yakınsayan paralel karma yöntem deseni (Creswell, 2011'den uyarlanmıştır).

Örnekleme

Bu çalışmada veri toplama süreci, 15'i BTÜ'de, 17'si Türkiye'deki çeşitli üniversitelerde görev yapan ve yabancı dil olarak İngilizce öğreten toplam 32 öğretim elemanı aracılığıyla gerçekleştirilecektir. Katılımcıların tamamı Türkiye'de İngilizce ile ilgili bir lisans bölümü (İngiliz Dili Eğitimi, İngiliz Dili ve Edebiyatı, İngiliz Dil Bilimi, Mütercim-Tercümanlık vs.) mezunu olup aynı anadili (Türkçe) konuşuyor olacaklardır. Projenin amacına uygun olarak, çalışmaya katılan puanlayıcıların tümü mesleki kariyerlerinde yazma becerisinin öğretimi ve yazma performansının değerlendirilmesi kapsamında farklı mesleki tecrübelere sahip olacaktır. Söz konusu örneklemin seçilmesindeki amaç, proje araştırmacısının örnekleme erişebilirliği ile birlikte, aynı kurum çatısı altında çalışan puanlayıcı ekibinin sağlayacağı verileri, kurum kültürü ve kurum beklentileri ile birlikte yorumlamaktır. Öte yandan, çalışma örnekleminin diğer yarısını oluşturacak olan katılımcıların farklı üniversitelerden seçilecek olması ise, puanlayıcıların örneklemden bağımsız bir şekilde kompozisyonları nasıl değerlendirdiğini görmek ve bulguları daha geniş bir pencereden ele almak olacaktır.

Aletler-Araç Gereç -Cihaz

Karma araştırma yöntemi kullanılan bu çalışmada nicel ve nitel veriler toplanacaktır. Öncelikle, her bir puanlayıcı tarafından 50 adet kompozisyona verilecek olan ve toplamda elde edilecek 1600 adet yazma performansı değerlendirme puanları nicel veriyi oluşturacaktır. Araştırmada analitik rubrik kullanılacaktır. Söz konusu rubrik, daha önce G Kuramı çerçevesinde değerlendirme rubriğin (analitik ve holistik) İngilizce kompozisyon puan değişkenliği üzerindeki etkisini araştıran bir çalışmadan (Han, 2013) uyarlanacaktır. Rubrik uyarlanırken, araştırmada yer alacak puanlayıcılardan farklı kalitedeki öğrenci kompozisyonlarını puanlayarak rubrik ile ilgili görüş alınacak ve gerekli değişiklikler her bir katılımcının katkıları ile yapılacaktır. Bu sayede, araştırmada yer alacak katılımcıların, etkinliğine ve kullanışlılığına inandığı bir puanlama anahtarı ile kompozisyonları değerlendirmeleri sağlanacaktır.

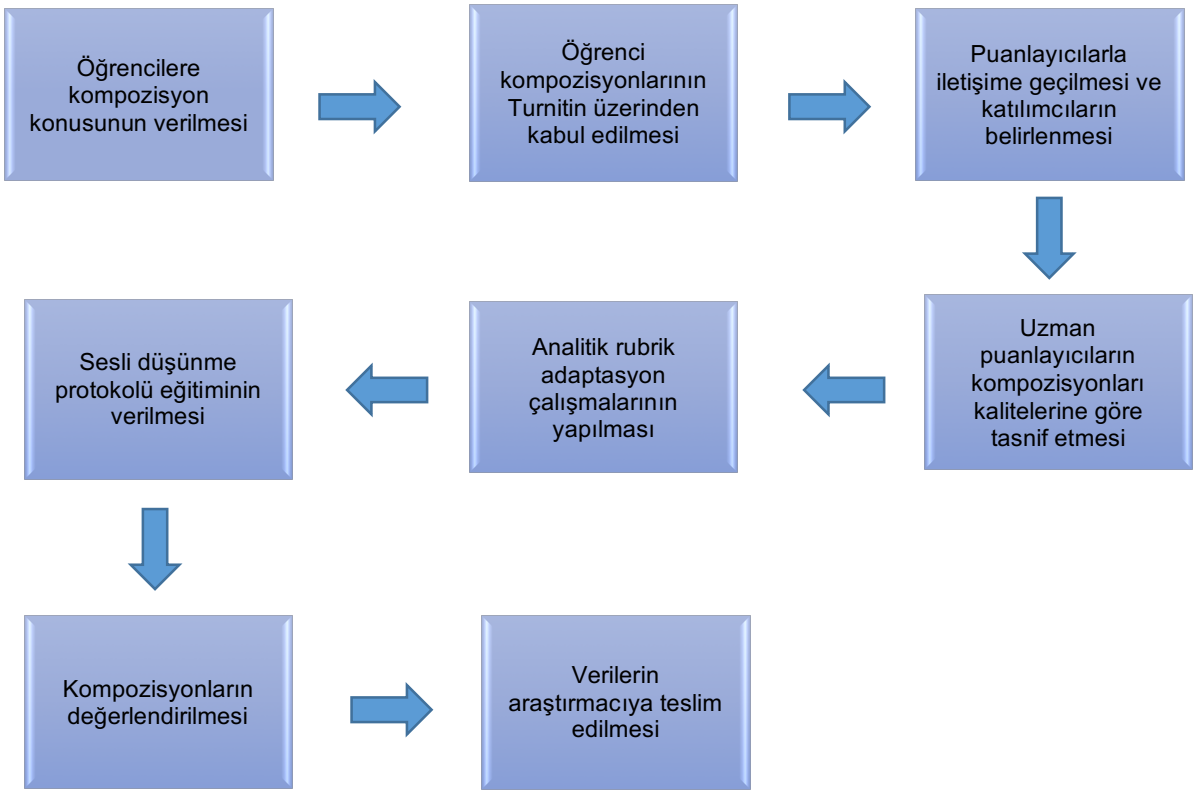
İkinci olarak, puanlayıcılardan kompozisyon değerlendirme işlemi sonrasında puanlayıcı profil formunu (rater's profile form) doldurmaları istenerek, özgeçmişleri ile ilgili veri toplanacaktır. Bu bilgiler ışığında puanlayıcılar mesleki tecrübelerine, cinsiyetlerine ve eğitim geçmişlerine göre gruplandırılarak nitel ve nicel veriler yorumlanacaktır.

Son olarak, puanlayıcının bilişsel yapısını ve karar verme süreçlerini yansıtacak olan nitel veri ise sesli düşünme protokolleri ile toplanacaktır. Sesli düşünme protokolleri hem birinci dilde (Huot, 1993; Wolfe, Kao ve Ranney, 1998) hem de ikici dilde (Barkaoui, 2007, 2010b; Cumming, Kantor ve Powers, 2002; Lumley, 2005) yazma performansı değerlendirme araştırmalarında kullanılmaktadır. Sesli düşünme protokolleri ile veriler toplanmadan önce, puanlayıcılar kompozisyon değerlendirme aşamasında düşüncelerini ve duygularını nasıl dışa vuracaklarıyla ilgili bir eğitim alacaklardır. Ayrıca, puanlayıcılara sesli düşünme protokollerinde dikkat edilmesi gereken hususlar ve izlemesi gereken adımlarla ilgili bir talimat listesi (instructions for think-aloud protocols; Cumming, Kantor ve Powers, 2001) verilecektir. Puanlayıcılar, kompozisyon değerlendirme sürecindeki düşüncelerini ve deneyimlerini ses kayıt cihazına kaydedeceklerdir. Daha sonra bu kayıtlar, araştırma ekibi tarafından yazılı metine dökülecektir. Sesli düşünme protokollerinden elde edilen veriler, Cumming ve diğerleri (2002) tarafından geliştirilmiş veri kodlama şeması (data coding scheme) kullanılarak analiz edilecektir.

Veri toplama

Proje araştırmacısı hem BTÜ hem de diğer üniversitelerde görev yapan öğretim elemanları ile irtibata geçmiş çalışmanın kapsamı ve amacı hakkında bilgilendirme yapmıştır. Yapılan bilgilendirme toplantıları BTÜ'de ve ulaşım kolaylığı olan üniversitelerde yüz yüze gerçekleşmiştir. Diğer katılımcılarla ise görüntülü konferans yöntemi ile iletişim sağlanmıştır. Daha sonra gönüllülük esasına bağlı kalarak araştırmada yer alacak katılımcılar belirlenmiştir. Araştırmada kullanılacak öğrenci kompozisyonları ÇOMÜ Eğitim Fakültesi Yabancı Diller Eğitimi Bölümü İngiliz Dili Eğitimi Anabilim Dalı'nda proje yürütücüsü tarafından verilmiş olan İleri Okuma ve Yazma Becerileri Dersi'ne kayıtlı öğrencilerden 2015-2016 akademik yılı bahar yarıyılında toplanmıştır. Öğrenci kompozisyonları, ortalama olarak 500 ile 700 kelime uzunluğundadır ve olası intihal olaylarını engellemek amacıyla intihal programı üzerinden kabul edilmiştir. Proje kapsamında toplanmış olan kompozisyonlar bağımsız 3 değerlendirmeci tarafından yüksek, orta ve düşük kalite olmak üzere üç başlık altında tasnif edilecektir. Bu şekilde orta kalitede yer alan kompozisyonlar tespit edilerek araştırmada kullanılmayacaktır. Uzman puanlayıcıların yüksek ve düşük kalite anlamında hemfikir oldukları kompozisyonlar çalışmaya dahil edilecektir. Nitekim puanlayıcılara kompozisyonların kalite bilgisi verilmeyecek ve öğrenci ödevleri rastgele kodlanacaktır. Takip eden süreçte proje ekibi, araştırmada yer alacak puanlayıcıların katkılarıyla çalışmada kullanılacak analitik ölçeğin adaptasyon çalışmalarını yapacak ve sonrasında pilot çalışma ile ölçeğin güvenirlik ve geçerlilik analizleri tamamlanacaktır. Anket çalışmasını takip eden süreçte katılımcı puanlayıcılara sesli düşünme protokolleri ile ilgili teorik ve uygulamalı bir eğitim verilecek ve kompozisyon değerlendirme aşamasında ihtiyaç halinde başvurmaları için kendilerine yazılı olarak sesli düşünme protokolü yönergesi temin edilecektir. Katılımcıların çalıştığı kurumlardaki eğitim-öğretim faaliyetlerinin tamamlanmasını takip eden periyotta, her bir puanlayıcıya 50 adet kompozisyon, 50 kopya analitik rubrik, sesli düşünme protokolü yönergesi, puanlayıcı profil formu ve ses kayıt cihazını içeren bir dosya temin edilecektir. Katılımcılardan 2 ay içerisinde kendilerine verilen kompozisyonları analitik ölçeği kullanarak puanlamaları ve kompozisyonların 16 tanesini

değerlendirirken sesli düşünmeleri ve bu süreci ses kayıt cihazına kaydetmeleri istenecektir. Puanlayıcılar, değerlendirme işlemini tamamladıktan sonra puanlayıcı profili formunu doldurarak, kendilerine teslim edilmiş bütün ekipman ve verileri araştırmacıya elden veya posta yoluyla teslim edecektir.



Şekil 2. Veri toplama süreci.

Veri Analizi

Araştırmada toplanan verinin analiz edilmesinde Statistical Package for Social Sciences (SPPS 20.0) ve Generalizability of Varience (GENOVA) programlarından yararlanılacaktır. Bu doğrultuda tanımlayıcı istatistik, bağımsız değişkenli *t*-test gibi analizlerle birlikte, genellenebilirlik kuramına bağlı olarak genellenebilirlik (G) ve karar (K) çalışmalarından yararlanılacaktır. Tecrübeli ve tecrübesiz puanlayıcıların öğrenci kompozisyonlarına verdiği puanların ortalamaları ve standart sapmaları tanımlayıcı istatistik ile saptanacaktır. Her iki puanlayıcı grubunu karşılaştırma amacıyla bağımsız değişkenli *t*-test kullanılacaktır. Genellenebilirlik kuramı çerçevesinde yapılacak olan analizlerle de değerlendirme puanlarının göstereceği değişkenlik kaynaklarının (facet) büyüklükleri ve etkileşim oranları belirlenecektir. Katılımcılardan toplanmış olan veriye araştırma grubu haricinde erişime izin verilmeyecektir. Bulgular rapor edilirken katılımcıların kimliklerini açığa çıkarabilecek herhangi bir bilgi verilmeyecektir.

Kısıtlama ve Sınırlamalar

Bu projede yabancı dil olarak İngilizce yazma performansı değerlendirme süreci araştırılacaktır. Bu bağlamda elde edilecek bulgular, hem üniversiteler hem de ÖSYM tarafından yapılan İngilizce dışındaki yabancı dillerdeki kompozisyon sınavlarının

puanlama süreci ile ilişkilendirilemez. Nitekim diğer yabancı diller için replike araştırmalar yapılabilir.

B Planı ve Önlemler

Veri toplama sürecinde karşılaşılabilecek olası sorunlar belirlenmeye çalışılmış ve bunlarla ilgili şu tedbirler alınmıştır. Projedeki veri toplama sürecinde kullanılacak kompozisyonlar proje yürütücüsünün İleri Okuma ve Yazma Becerileri Dersi'nde temin edilmiştir. Bu yüzden, söz konusu dersin, proje yürütücüsü tarafından 2015-2016 akademik yılı bahar yarı yılında da vermeye devam etmesi sağlanmıştır. Yabancı Diller Eğitimi Bölüm Başkanlığı bu proje önerisiyle ilgili bilgilendirilmiş ve açılan ders için proje yürütücüsünün görevlendirilmesi talep edilmiştir. Puanlayıcıların eğitim-öğretim faaliyetlerinin veri toplama sürecini olumsuz etkilememesi adına, puanlayıcalardan gelecek olan verinin 2015-2016 akademik yılının bitimini takiben toplanması planlanmıştır. Katılımcıların çalışmaya başladıktan sonra araştırmadan çekilme ihtimali göz önünde bulundurularak yedek katılımcı listesi oluşturulmuştur ve kendilerinin yedek katılımcı olarak çalışmaya katılımları ile ilgili olarak onayları alınmıştır.

Beklenmedik Durumlara Müdahale

Risk 1: Kompozisyonlara kalite tasnifini yapacak bağımsız değerlendirmecilerin herhangi birinin görevden çekilmesi

B Planı: Alanında uzman yedek değerlendirmecilerden onay alınmıştır.

Risk 2: Araştırmada kullanılacak analitik ölçeğin adaptasyonu

B Planı: Sürecin her aşamasında uzman görüşü alınarak gerekli görüldüğü taktirde ilgili maddeler üzerinde adaptasyon yapılacaktır.

Risk 3: Katılımcılara anket kullanımı ve sesli düşünme protokolleri uygulama esasları ile ilgili eğitimin verilmesi

B Planı: Puanlayıcıların bu tür veri toplama yöntemine aşina olması için kuramsal ve uygulama eğitimi yapılacaktır. Ayrıca, veri toplama araçlarının bulunduğu paketin içine sesli düşünme protokolü yönergesi ve ses-kayıt cihazına örnek bir ses dosyası eklenerek puanlayıcıların söz konusu veri toplama yöntemini içselleştirmeleri sağlanacaktır.

Risk 4: Öğrenci kompozisyonlarının değerlendirilmesi

B Planı: Değerlendirme işleminin zamanında tamamlanabilmesi için araştırmacı puanlayıcılarla belirli periyotlarda iletişime geçecektir.

Risk 5: SPSS veri girişi

B Planı: Veri girişinde herhangi bir hatanın yapılmaması için, girilen veri yürütücü denetiminde araştırmacı tarafından kontrol edilerek girilecektir.

Risk 6: Sesli düşünme protokolleri ses kayıtlarının yazılı metine dökülmesi

B Planı: Herhangi bir sebeple oluşabilecek veri kaybı ihtimaline karşın, ses kayıtlarının transkrisiyonları eş zamanlı olarak yedeklenecektir.

Risk 7: Sesli düşünme protokollerinin içerik analizi (kodlanması)

B Planı: Barkaoui (2007) farklı türde ölçek kullanımının puanlama süreci ve puanlayıcıların karar verme stratejileri üzerindeki etkisini incelemiştir ve bu çalışmada, sesli düşünme protokolleri ile elde edilen verinin %15'i bağımsız bir değerlendirmeci tarafından kodlanarak nitel veri analizinin güvenirliği kontrol edilmiştir. Benzer şekilde,bu araştırmada da sesli düşünme protokolleri ile elde edilecek verinin %15'i (77 kayıt) bağımsız bir araştırmacı tarafından kodlanarak araştırmacı tarafından yapılan kodlamaların güvenirlik kontrolü yapılacaktır.

Risk 8: Veri analizi

B Planı: Veri analizinde herhangi bir hatadan kaçınmak için, veri analizi hem yürütücü hem de araştırmacı tarafından ve sonuçlar karşılaştırılarak kontrol edilecektir.

Risk 9: Bulguların literatürle ilişkilendirilerek yorumlanması

B Planı: Bulguların yorumlanmasında herhangi bir hatadan kaçınmak için, analiz tabloları hem yürütücü hem de araştırmacı tarafından yorumlanacak ve yorumlar karşılaştırılarak kontrol edilecektir.

Çalışmanın Olası Etkileri

Çalışma başarıyla gerçekleştirildiği takdirde ulaşılması beklenen etkiler aşağıdaki gibidir.

Çalışmanın bulgularının 2017 yılında 16. kez düzenlenecek olan çalışma konusuyla ilgili en prestijli organizasyonlardan birisi olarak kabul edilen İkinci Dilde Yazma Sempozyumu'nda (Symposium on Second Language Writing) sunulması planlanmaktadır. Ayrıca, araştırma bulgularının Elsevier tarafından yayınlanmakta olan Journal of Second Language Writing ve SAGE tarafından yayınlanmakta olan Language Testing dergilerinde yayınlanmak üzere gönderilmesi planlanmaktadır.

Projenin sonuçlanmasıyla, üniversiteler ve ÖSYM tarafından kullanılabilecek yazma performansı analitik değerlendirme ölçeği geliştirilecektir. Elde edilen bulgular neticesinde, özgün puanlayıcı eğitimi içeriği oluşturulup, TÜBİTAK desteği ile yükseköğretim kurumlarında görev yapan ve yabancı dil olarak İngilizce öğreten öğretim elemanlarına eğitim verilmesi düşünülmektedir. Bu sayede kompozisyon değerlendirme sürecinin standartlaşmasına katkıda bulunulacak ve puanlayıcı güvenirliği sağlanmış olacaktır.

Bu araştırma, danışmanlığı proje yürütücüsü tarafından yürütülen bir doktora çalışmasıdır. İlgili doktora öğrencisi de bu projede araştırmacı olarak görev yapacaktır. Bu açıdan bu proje, Anabilim Dalı'mızda yürütülmekte olan yüksek lisans ve doktora programında, proje konusunu daha ileriye götürebilmeye istekli lisansüstü öğrencilerini araştırma yapmaya teşvik edecektir. Bu nedenle, çalışmanın bulgularının söz konusu öğrencilerin tez önerilerinin şekillenmesine de katkıda bulunacağı düşünülmektedir. Bu anlamda gelecekteki projelerde karmaşık doğalarından ötürü tercih edilmeyen istatistiksel yöntemlerle yazma becerisinin değerlendirilmesindeki sorunlar farklı açılardan ele alınacaktır.

Kaynakça

Baba, K. 2009. "Aspects of lexical proficiency in writing summaries in a foreign language". Journal of Second Language Writing, 18, 191-208. doi:10.1016/j.jslw.2009.05.003

Bachman, L. F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Barkaoui, K. 2007. "Rating scale impact on EFL essay marking: A mixed-method study". Assessing writing, 12(2), 86-107.

Barkaoui, K. (2008). Effects of Scoring Method and Rater Experience on ESL Essay Rating Processes and Outcomes. Yayınlanmamış doktora tezi, University of Toronto, Kanada.

Barkaoui, K. 2010a. "Explaining ESL essay holistic scores: a multilevel modeling approach". Language Testing, 27(4), 515-535.

Barkaoui, K. 2010b. "Do ESL essays raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study". TESOL Quarterly, 44(1), 31-57.

Barkaoui, K. 2010c. "Variability in ESL essay rating processes: The role of the rating scale and rater experience". Language Assessment Quarterly, 7(1), 54-74.

Barkaoui, K. 2011. "Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity". Language Testing, 28(1), 51-75.

Breland, H. M. 1983. The Direct Assessment of Writing Skill: A Measurement Review (ETS Research Report No: 86-9). Princeton, NJ: Educational Testing Service.

Brown, J. D. 1991. "Do English and ESL faculties rate writing samples differently?" TESOL Quarterly, 25(4), 587-603.

Cresswell, J. W. (2011). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.). New Delhi, India: PHI Learning Private Ltd.

Cumming, A. 1990. "Expertise in evaluating second language composition". Language Testing, 7(1), 31-51.

Cumming, A., Kantor, R., ve Powers, D. 2001. Scoring TOEFL Essays and TOEFL 2000 Prototype Tasks: An Investigation into Raters' Decision Making and Development of a Preliminary Analytic Framework (TOEFL Monograph Series, Report No: 22).Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., ve Powers, D. 2002. "Decision making while rating ESL/EFL writing tasks: A descriptive framework". Modern Language Journal, 86(1), 67-96.

Dörnyei, Z. 2007. Research Methods in Applied Linguistics: Quantitative, Qualitative and Mixed Methodologies. Oxford: Oxford University Press.

Eckes, T. (2012). "Operational rater types in writing assessment: Linking rater cognition to rater behavior". Language Assessment Quarterly, 9(3), 270-292.

Elorbany, R., ve Huang, J. 2012. "Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach". Language and Communication Quarterly, 1, 2-24.

Hamp-Lyons, L. 1996. "The challenges of second language writing assessment". Assessment of Writing: Policies, Politics, Practice. Editörler: White, E., Lutz, W., Kamusikiri, S. New York: Modern Language Association.

Han, T. 2013. The Impact of Rating Methods and Rater Training on the Variability and Reliability of EFL Students' Classroom-based Writing Assessments in Turkish Universities: An Investigation of Problems and Solutions. Yayınlanmamış doktora tezi, Atatürk Üniversitesi..

Homburg, T. J. 1984. “Holistic evaluation of ESL composition: can it be validated objectively”? TESOL Quarterly, 18(1), 87-108.

Huang, J. (2007). Examining the Fairness of Rating ESL Students’ Writing on Large-scale Assessments. Yayınlanmamış doktora tezi, Queen’s University, Kanada.

Huang, J. 2008. “How accurate are ESL students’ holistic writing scores on large-scale assessments?—A generalizability theory approach”. Assessing Writing, 13(3), 201-218.

Huang, J. 2009. “Factors affecting the assessment of ESL students’ writing”. International Journal of Applied Educational Studies, 5(1), 1-17.

Huang, J. 2011. “Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments”. TESOL Journal, 2(4), 423-443.

Huang, J. 2012. “Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment”. Assessing Writing, 17, 123-139.

Huang, J., ve Foote, C. J. 2010. “Grading between lines: What really impacts professors’ holistic evaluation of ESL graduate student writing”? Language Assessment Quarterly, 7(3), 219-233.

Huang, J., ve Han, T. (2013). “Holistic or analytic –A dilemma for professors to score EFL essays”? Leadership and Policy Quarterly, 2(1), 1-18.

Huang, J., Han, T., Tavano, H., ve Hairston, L. 2014. “Using generalizability theory to examine the impact of essay quality on rating variability and reliability of ESOL writing”. Empirical Quantitative Research in Social Sciences: Examining Significant Differences and Relationships. Editörler: Huang, J., Han, T.. New York: Untested Ideas Research Center.

Huot, B. A. 1990. “Reliability, validity and holistic scoring: What we know and what we need to know”. College Composition and Communication, 41(2), 201-213.

Huot, B. A. 1993. “The influence of holistic scoring procedures on reading and rating student essays”. Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations. Editörler: Williamson, M. M., Huot, B. A. Cresskill, NJ: Hampton Press.

Lim, G. S. 2011. “The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters”. Language Testing, 28(4), 543-560.

Linn, R. L., ve Burton, E. 1994. “Performance-based assessments: Implications of task specificity”. Educational Measurement: Issues and Practice, 13(1), 5-8, 15.

Lumley, T. 2005. Assessing Second Language Writing: The Rater’s Perspective. New York: Peter Lang.

McColly, W. (1970). “What Does Educational Research Say About the Judging of Writing Ability?” The Journal of Educational Research, 64(4), 148-156.

Reid, J., ve O’Brien, M. 1981. “The application of holistic grading in an ESL writing program” [Proceeding]. Annual Convention of Teachers of English to Speakers of Other Languages. Detroit, MI. (ERIC No. ED 221 044).

Rinnert, C., ve Kobayashi, H. 2001. “Differing perceptions of EFL writing among readers in Japan”. The Modern Language Journal, 85, 189-209.

Shohamy, E., Gordon, C. M., ve Kraemer, R. 1992. “The effect of raters’ background and training on the reliability of direct writing tests”. The Modern Language Journal, 76, 27-33.

Song, B., ve Caruso, I. 1996. “Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students”? Journal of Second Language Writing, 5, 163-182.

Stalnaker, J. M., ve Stalnaker, R. C. 1934. “Reliable reading of essay rests”. The School Review, 42(8), 599-605.

Sweedler-Brown, C. O. 1985. “The influence of training and experience on holistic essay evaluation”. English Journal, 74, 49-55.
Weigle, S. C. (1994). Effects of training on raters of ESL compositions. Language Testing, 11,197-223.
Weigle, S. C. 1999. “Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches”. Assessing Writing, 6, 145-178.
Weigle, S. C. 2002. Assessing Writing. Cambridge: Cambridge University Press.
Wolfe E. W., Kao, C. & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. Written Communication, 15, 465–492.